DOCTOR OF SCIENCES

Stochastic processes on temporal networks
spreading strategies and emergence of memory

Gueuning, Martin

Award date:
2019

Awarding institution:
University of Namur

UNIVERSITÉ DE NAMUR **UNIVERSITÉ CATHOLIQUE DE LOUVAIN**

FACULTÉ DES SCIENCES FACULTÉ DES SCIENCES

DÉPARTEMENT DE MATHÉMATIQUE ICTEAM

Stochastic processes on temporal networks: spreading strategies and emergence of memory

Thèse présentée par
Martin Gueuning
pour l'obtention du grade
de Docteur en Sciences

Composition du Jury :

Timoteo CARLETTI
Jean-Charles DELVENNE (Promoteur)
Julien HENDRICKX
Renaud LAMBIOTTE (Promoteur)
Anne LEMAITRE (Président du Jury)
Luis E C ROCHA

9 Septembre 2019

Graphisme de couverture : © Presses universitaires de Namur

Rempart de la Vierge, 13
B-5000 Namur (Belgique)

Imprimé en Belgique

ISBN : 978-2-39029-064-3
Dépôt légal : D/2019/1881/18

Université de Namur
Faculté des Sciences
rue de Bruxelles, 61, B-5000 Namur (Belgique)

Processus stochastiques sur réseaux temporels : stratégies de diffusion et émergence de mémoire.

par Martin Gueuning

Résumé : La théorie des réseaux permet de modéliser des sytèmes issus de domaines variés. L'augmentation récente du volume de données empiriques a permis de prendre en compte des caractéristiques plus réalistes dans la modélisation des réseaux. En particulier, l'étude des détails concernant la temporalité des interactions entre les agents a mis en avant la nature non-markovienne de leur comportement. Dans ce contexte, ce travail étudie des processus stochastiques sur des réseaux temporels et se décompose en deux parties. La première étudie l'impact d'un comportement non-markovien sur une marche aléatoire effectuée sur un réseau temporel. Nous mettons en évidence l'émergence de mémoire dans la trajectoire du marcheur due aux comportements "bursty" et à la présence de cycles courts. La seconde partie de ce travail s'intéresse à la diffusion d'information dans les réseaux sociaux. Nous illustrons comment la série temporelle de retweets d'une cascade fournit des informations sur la manière dont le message s'est propagé dans le réseau. Sur base de ces résultats, nous développons l'algorithme *SmartInf* dont l'objectif est de fournir une liste d'utilisateurs à viser simultanément afin de maximiser le nombre de partages d'un message initial. *SmartInf* se base sur les propriétés temporelles des cascades et sur une connaissance partielle de la structure du réseau, contrairement aux méthodes standards qui nécessitent la connaissance totale du réseau qui est généralement coûteuse .

Stochastic processes on temporal networks: spreading strategies and emergence of memory

by Martin Gueuning

Abstract: Network theory provides a framework to model interacting systems from many different fields. The recent increase in the availability of empirical data has allowed the research community to take into account more realistic characteristics of the networks. In particular, the details of the timing of the interactions have highlighted the non-Markovian nature of the agents in many systems. This work lies in the context of stochastic processes on temporal network and is twofold. The first part of this work aims at studying the impact of non-Markovian activities on Random Walks taking place on temporal networks. We show that memory in the trajectory of the random walker naturally emerges due to bursty behaviours and the presence of short cycles. The second part of this work aims at designing efficient spreading strategies on temporal networks. In particular, we show that the temporal sequence of a cascade of retweets may provide an insight about the way it spread on the network. We exploit these findings by developing the *SmartInf* algorithm that provides a list of users to target simultaneously, in order to maximize the final share of a message. Importantly, *SmartInf* relies on temporal patterns of cascades and on local structural information, in opposition to standard methods that rely on the typically costly global structure.

Thèse de doctorat en Sciences Mathématiques (Ph.D. thesis in Mathematics)
Date: 2019
Département de Mathématique
Promoteurs (Advisors): Jean-Charles DELVENNE, Renaud LAMBIOTTE

Remerciements

Ça y est ! Après tout ce temps passé sur cette thèse, cela me semble un peu irréel d'en arriver au bout. Il y a une foule de gens qui ont contribué à rendre cela possible.

En premier lieu, j'aimerais remercier mes promoteurs Jean-Charles Delvenne et Renaud Lambiotte pour avoir accepté de me superviser tout au long de la thèse. Ils m'ont permis de développer mon autonomie tout en gardant un regard critique sur mon travail et en y apportant des commentaires de grande valeur. J'ai toujours été impressionné par leur qualité d'écriture et leur large vision de la recherche. Merci aussi d'avoir rendu possibles les différentes collaborations que j'ai pu développer durant ma thèse.

Merci à Anne Lemaitre d'avoir accepté de présider mon jury. Je serai toujours reconnaissant pour l'aide qu'elle m'a apportée lors de mon Master. Merci à Timoteo Carletti, Julien Hendrickx et Luis E C Rocha d'avoir accepté de faire partie de mon comité d'accompagnement et de mon jury. Les commentaires de mon jury m'ont permis d'améliorer substantiellement la qualité et la clarté de cette thèse.

Merci à Julien Petit et Sibo Cheng pour leur motivation dans les collaborations que nous avons développées, c'est toujours plus motivant de travailler ensemble. Thank you to Ayan Kumar Bhowmick and Bivas Mitra for their collaboration and their hospitality at the Indian Institute of Technology of Kharagpur.

Je souhaiterais également remercier le PAI Dysco et l'Université de Namur pour m'avoir financé durant ma recherche.

Je remercie mes collègues de bureau (ou assimilés) qui ont transformé le bureau en un lieu plus qu'agréable à fréquenter : Jéméry en particulier pour sa patience envers mes goûts musicaux (les murs de Grenoble en tremblent encore) et pour m'avoir fait déculpabiliser par rapport à mon organisation, Jon et sa poésie effrenée, Alexis Roi d'Italie, Watson et ses citations, et enfin Nico qui repousse sans cesse les frontières de l'humour. Merci aussi à Allan et David pour m'avoir accueilli à l'Euler pendant 4 ans.

Merci à mes collègues de couloir ou de repas à l'Arsenal qui ont égayé mes nombreuses pauses repas, belote et bulot (Pauline, Manon, Delphine, Eve, Mara, Ambi, Marie, Morgane, Candy, Julien, Francois, Arnaud, Joanna et tous les autres).

Merci à Benjou et Hélène pour toutes ces escapades et ces soirées, vous êtes les plus crochus, tout simplement.

Enfin, merci à ma famille pour son soutien inconditionnel.

Et finalement, Juliette, merci pour tout ! Pour les milliers de fois où tu m'as remonté le moral, pour ta patience quand je te parle de la magie des mathématiques, pour tous ces moments passés ensemble, ceux à venir et pour tout le reste.

Contents

Introduction **1**

I Preliminaries 7

1 Theoretical Background **9**
1.1 Fundamentals of the network theory 9
1.2 Complex networks 12
1.3 Stochastic processess 12
1.4 Burstiness 14
1.5 Random Walks on temporal networks 16
1.5.1 Active node-centric Random Walk 16
1.5.2 Active edge-centric Random Walk 19
1.5.3 Passive edge-centric Random Walk and the bus paradox . . . 20
1.5.4 Extensions of edge/node-centric Random Walk 23
1.5.5 Random Walks on multilayer networks 24
1.6 Spreading models 26
1.6.1 Compartmental models 27
1.6.2 Models of cascade diffusion 28
1.6.3 Identification of influential users 30
1.7 Data collection on Twitter 31

II Random Walks on temporal networks: emergence of memory 33

2 Backtracking and Mixing Rate **35**
2.1 Introduction 36
2.2 Passive edge-centric Random Walk on temporal network 36
2.2.1 Bias on the probability of backtracking 36
2.2.2 Impact of backtracking on the mixing rate of the Random Walk 41

2.3 Discussion 44

3 Rock-Paper-Scissors Dynamics 47
3.1 Introduction 48
3.2 Active edge-centric Random Walk on multiplex temporal networks . . 49
3.2.1 Emergence of biased paths 49
3.2.2 Basic properties of precedence 50
3.3 Impact on coverage 54
3.4 Discussion 56

4 Random Walk with lasting edges 59
4.1 Introduction 60
4.2 Random Walk with lasting edges 60
4.2.1 Model description 60
4.3 Discussion on the different timescales 61
4.4 Dynamics on a Directed Acyclic Graph 65
4.4.1 Master equation on a DAG 65
4.4.2 Transition density on a DAG 67
4.4.3 Limiting cases with exponential distributions on a DAG . . . 69
4.5 Dynamics on directed networks with cycles 70
4.5.1 Generalized master equation with correction for cycles of length two 70
4.5.2 Transition density with correction for cycles of length two . . 72
4.6 Discussion 76

III Spreading strategies on temporal networks 79

5 Imperfect spreading on temporal networks 81
5.1 Introduction 82
5.2 Imperfect spreading on temporal networks 82
5.2.1 Estimating the inter-success time 82
5.2.2 Epidemic threshold and spreading efficiency 85
5.3 Discussion 90

6 Temporal sequence of retweets reveals cascade migration 91
6.1 Introduction 92
6.2 Datasets 93
6.3 Inter-retweet intervals and spreading dynamics 94
6.3.1 Sorting cascades based on inter-retweet intervals 94
6.3.2 Cascade diffusion across diffusion localities 96
6.3.3 Co-occurrence between early peaks and flushes 97
6.4 Analytical model for cascade spreading across diffusion localities . . 98
6.4.1 Modeling cascade spreading in a single diffusion locality . . . 99
6.4.2 Modeling cascade migration across multiple localities 101
6.5 Empirical validation 104

6.5.1 Cascade migration after the first peak 104
6.5.2 Locality saturation and inter-retweet intervals 106
6.5.3 Remarks on anchor nodes 106
6.6 Discussion 107

7 Detecting influential nodes from temporal sequences 109
7.1 Introduction 110
7.2 The SmartInf algorithm 111
7.2.1 Problem statement 111
7.2.2 SmartInf algorithm description 112
7.2.3 Computational complexity 113
7.2.4 Importance of the refining step 113
7.2.5 Standard centrality measures 114
7.3 Evaluation of SmartInf performance on empirical data 115
7.3.1 Baseline algorithms 115
7.3.2 Twitter-specific metrics 116
7.3.3 Epidemic simulation 118
7.3.4 Quality of influential nodes obtained by Smartinf 120
7.3.5 Robustness of SmartInf 121
7.4 Evaluation of SmartInf performance on synthetic data 122
7.4.1 Synthetic setup 122
7.4.2 Experimental observations 125
7.5 Discussion 126

Discussion 129

Bibliography 133

Introduction

Context

Network theory allows one to model interconnected or interacting complex systems from many different fields. Important examples include the Internet, (online) social networks, airline routes, but also a wide-range of biological networks, such as food webs, the brain or protein interaction networks. An interacting entity is represented by a node, and the interactions between two entities by an edge. The study of networks has emerged in the last decades as one of the fundamental building blocks in the wider study of complex systems. One of the main reasons of its success is the possibility to analyze systems of very different natures within a single framework , notably helping in the quest of finding universal laws.

Initially, graph theorists focused on small or regular networks then turned to larger random graphs with simple characteristics. The sudden increase of available data allowed to turn to more realistic models of networks, which intrinsically display some organization. However, networks are often dynamical rather than static entities as nodes and edges may be created or destroyed over time. Alternatively, only varying subgraphs of the underlying static network are available at different periods of time. Network theory has naturally evolved to take into account the dynamic nature of the interactions.

The exploration of empirical data has revealed that the nature of the interactions is typically not Markovian. The probability distributions associated to the times between two consecutive actions of an entity or between two consecutive interactions between two entities are not exponential and tend to be heavier tailed, which means that large values of these times are more likely to be observed than it would if the times where following an exponential distribution with equal mean. The non-Markovian nature of the processes is due to a variety of different reasons, such as the bursty behaviour of the agents, the correlations between consecutive interactions, the ordering of the interactions or the presence of memory. The existence of inherent periodic cycles also leads to non-stationary processes.

An aggregated static network may provide an inaccurate representation of the interacting system it models, which may be tackled using a temporal network framework. For instance, aggregated networks may contain non-realistic paths. Let us assume a static network aggregating only two interactions: one between A and B at time t_1 and another one between B and C at time $t_2 > t_1$. Then, the static network contains a path from C to A but it is actually not possible to move from C to A. Temporal networks on the contrary allow one to consider only time-preserving paths. Moreover, temporal networks also allow one to distinguish between the shortest paths in terms of length and the ones in terms of duration. Indeed, these paths may significantly differ because the interactions times between the entities may be of distinct timescales. Therefore, the extension from static to temporal networks provides more accurate models but also allows one to study the impact of properties which are impossible to be taken into account in a static network framework.

Moreover, networks may serve as media of diffusion for entities that are independent from the network dynamics. For instance, a virus may propagate over a social network where the interactions between people are independent from the existence of the virus. Thus, there may exist an interplay between the dynamics of a network itself and the entity that is propagating on it. The understanding of the impacts of the temporal properties of a network on the dynamics of the propagating entity is of paramount importance for the understanding of many stochastic processes taking place on the network, and furthermore to take appropriate actions for decision makers in various contexts such as epidemic prevention, news diffusion or urban planning, to name a few. The involved processes are stochastic because of the network and because of their intrinsic dynamics. In this thesis, we will focus on two distinct families of stochastic processes: diffusion processes, in particular Random Walks, and spreading processes.

Random Walks form a standard family of stochastic processes on networks, where an entity is assumed to jump between the nodes of the network following some specific decision rules which specify the nature of the walk. Despite the apparent simplicity of the process, Random Walks provide a better understanding of diffusion processes on complex temporal networks, and their usage spreads from simplified diffusion model on networks to center-piece of numerous algorithms designed to detect central nodes and community structures. In particular, the study of such dynamical processes allows to highlight the effects on the dynamics of stochastic processes on networks that are due to their topology or to their temporal characteristics. For instance, one recent debate in the literature we will discuss in section 1.5.1 was to determine whether bursty activity result in a speed up or in a slow-down of the diffusion, and when the topological properties of the network impact the dynamics more than the temporal ones.

A second kind of stochastic processes consists in spreading models where an entity is assumed to spread over the network by leaving copies of itself at each transmission. Spreading models are designed to study the propagation of diseases or information over a population. Leveraging on the understanding of the diffusing dynamics on temporal social networks, one may design efficient strategies to prevent a pandemic or

to diffuse a message faster. The emergence of online social networks such as Twitter has allowed to collect large samples of information cascades publicly available when users exhibit bursty behaviours. However, the 2014 Facebook-Cambridge Analytica scandal raised concerns in the society about the exploitation of personal data. One of the resulting challenge in terms of data analysis is to develop new tools that help the decision makers while being less intrusive than the standard methods in terms of personal data acquisition. A direction we follow in this thesis is the exploitation of the time series of specific public messages rather than the complete list of the interactions between all the users.

In this thesis, we consider both Random Walks and spreading models on temporal networks. First, we investigate theoretical models of Random Walks on temporal networks in order to highlight some effects which are induced by the non-Markovian nature of the diffusion process. In particular, we exhibit the emergence of memory and biases in the trajectory of the walker due to backtracking, non-vanishing edges and rock-paper-scissors effects. Then, we first theoretically study the trade-off between the quality and the volume of diffusion of a message when resources are limited and the users' behaviour is bursty. Finally, we switch from a model-driven to a data-driven approach and investigate datasets extracted from Twitter to exhibit how structural information may be deduced only from the timing of the retweets of a popular tweet. Based on these findings, we finally develop an algorithm that provides a list of preferential users to target in order to maximize the share of a given tweet.

Structure of the thesis

Part I: Preliminaries

In Chapter 1 we provide a short theoretical background in the context of Random Walks and spreading models on temporal networks. We formally introduce the concepts of burstiness, bus paradox, multilayers and the distinction between active and passive as well as edge-centric and node-centric Random Walks. Then we provide a brief state-of-the art of the models of cascade diffusion and of the existing methods for the identification of influential users for information diffusion. Finally we shortly introduce how to collect data on Twitter.

Part II: Random Walks on temporal networks: emergence of memory

In Chapter 2 we study the emergence of backtracking in the edge-centric Random Walk on temporal networks, an underlying effect that had been neglected in the literature. Backtracking is the tendency of the walker to jump back to the last visited node. We highlight the emergence of such backtracking bias effect for bursty walkers and quantify its impact on the mixing rate of the walker. We discuss the implications of our results on the elaboration of standard uncorrelated null models on temporal networks which neglect this emerging bias.

In Chapter 3 we consider an edge-centric Random Walk on a multiplex network where the distributions associated to the edges are identical for the edges belonging to the same layer but may differ between layers. We study the notion of precedence between layers and show the emergence of phenomena trapping the walker, either on a single layer or on an oriented cycle, even though edges activations are independent. We numerically illustrate the resulting slow-down on the coverage of the network.

In Chapter 4 we consider a Random Walk on a temporal network when the activations of the edges are not infinitesimal, in opposition to the classical simplifying assumption that the durations of the activation of the edges are negligible. In the context of non-vanishing edges, one needs to consider three timescales, associated to the walker, to the duration of the activation of an edge and to the time between two consecutive activations of an edge. We first consider the walk on a directed acyclic graph (DAG) and show that our general model reduces to the standard models in particular situations. We then extend it to graphs containing cycles and exhibit biases on the trajectory due to short cycles. We show that memory emerges even when all the processes are Markovian. Again, such memory may lead to the capture of the walker on a cycle. We develop a method to compute the memory corrections to the equation obtained for the DAG. We provide the explicit corrections for cycles of length two but the principle extends to larger cycles.

Part III: Spreading strategies on temporal networks

In Chapter 5 we focus on the impact of a bursty behaviour on the speed of the diffusion of a message in the context of an imperfect spreading. We consider that the information may be shared at each interaction with some given probability p, implying that several contacts may be required for the information to spread between two users. We explore the situation where the spreader has to choose between improving the quality of the message (that is the probability p of success of transmission over one interaction) or the rate of the contacts. We show that the spreading between two individuals is more efficient in terms of speed if the spreader decides to focus on the quality of the message rather than on the contact rates, or in other terms, that it is more efficient to charm than to spam.

In Chapter 6 we study datasets of cascades that spread on Twitter. We focus on the inter-retweet intervals of a cascade, that is the time between any two retweets of a given seed tweet. We classify the cascades into two types, depending on the occurrence of at least one early peak in the inter-retweet intervals of the cascades. Then we show that this peak reveals a phenomenon of saturation of the tweet on a locality of the network followed with the consecutive migration of the tweet to a new locality. Finally, we provide a simple analytical model that predicts such co-occurrence and validate it on empirical data. Therefore, we show that temporal patterns allow to gain insight about how the tweet spread on the network without any knowledge of its topology.

In Chapter 7 we directly exploit the findings of Chapter 6 in order to detect influential users. We consider the problem of multi-seeds targeting, which consists in providing a set of users to target in order to maximize the final spread of a message. We provide the *SmartInf* algorithm to tackle this problem. *SmartInf* relies on the refinement of a list of individuals who tend to be active around the intermediate peaks of the inter-events of a cascade. Such target list may be established at a cheaper cost as only temporal information is required. The refinement step only requires the availability of some local structure of some specific users, in contrast with standard algorithms that require the costly knowledge of the structure of the global network. We show that *SmartInf* outperforms baseline algorithms based on Twitter-specific metrics and on numerical simulations. We conclude with the investigation of some key factors behind the performance of *SmartInf*.

We conclude this thesis with a final discussion about our results and provide some future research directions in the context of temporal networks.

Contributions

The results of the chapters 2 to 7 correspond to the following publications. I fully contributed to the results of the publications in which I am mentioned as first author and to the results of [Petit et al. 2018] that are presented in Chapter 4. I also contributed to all the results of [Bhowmick et al. 2017] and [Bhowmick et al. 2019] that are presented in this monograph except for the analysis and simulations on empirical data which I co-designed but did not code in Python.

Gueuning, M.; Lambiotte, R.; Delvenne, J.-C. Backtracking and Mixing Rate of Diffusion on Uncorrelated Temporal Networks. *Entropy*, 19, 542. (2017)

Gueuning, M.; Cheng, S.; Lambiotte, R.; Delvenne, J.-C. Rock-Paper-Scissors dynamics from random walks on temporal multiplex networks. *Journal of Complex Networks*, cnz027 (2019)

Petit, J.; Gueuning, M.; Carletti, T.; Lauwens, B.; Lambiotte, R. Random walk on temporal networks with lasting edges. *Phys. Rev. E* 98, 052307 (2018)

Gueuning, M.; Delvenne, J.-C.; Lambiotte, R. Imperfect spreading on temporal networks. *Eur. Phys. J. B* 88: 282 (2015)

Bhowmick, A. K.; Gueuning, M.; Delvenne, J.-C.; Lambiotte, R.; Mitra, B. Temporal pattern of (re) tweets reveal cascade migration. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2017)

Bhowmick, A. K.; Gueuning, M.; Delvenne, J.-C.; Lambiotte, R.; Mitra, B. Temporal Sequence of Retweets Help to Detect Influential Nodes in Social Networks. *IEEE Transactions on Computational Social Systems* (2019)

Part I

Preliminaries

Chapter 1

Theoretical Background

1.1 Fundamentals of the network theory

The concept of networks is broad and general [Newman 2010]. It describes how entities are connected to each other and applies to various fields, including technical infrastructures, economy, biology, ecology, sociology or (online) social media, to name a few. Each entity is represented by a node, and the possibility of an interaction between two entities is represented by an edge between the two corresponding nodes. Such edges may be directed our undirected. There exist several ways to encode the structure of a network of n nodes and m edges.

The adjacency matrix A is a $n \times n$ matrix that encodes the relationship between the nodes and whose elements are defined as

$$A_{ij} = \begin{cases} 2 & \text{if } i = j \text{ and there exists a loop connecting the node } u_i \text{ to itself,} \\ 1 & \text{if there exists an edge connecting the distinct nodes } u_i \text{ and } u_j, \\ 0 & \text{otherwise.} \end{cases}$$

If the network is weighted, one may replace the values 1 and 2 by the weights of the corresponding edges, which may be negative for signed networks. Note that there are two conventions for dealing with the loops, the second one encoding a value of 1 in the diagonal of A for the loops. However, we will always consider networks without loops in this monograph.

The adjacency matrix provides several information about the corresponding network. For instance, the k^{th} power of the matrix A provides the number A^k_{ij} of path of length k between any two nodes u_i and u_j. The study of the eigenvalues of A provides insights about the structural organization of the network and allows one to determine the relaxation time of a diffusive process taking place on the network [Lovász 1993].

The line graph associated to a network allows to explicitly take into account the paths of length two. In such a graph, a node corresponds to an edge of the initial network, and an edge between two nodes exists if the two corresponding edges are incident in the initial network, that is if there exists a path of length two formed by these two edges.

The incidence matrix K is a $n \times m$ matrix that encodes the relationship between the nodes and the edges. Each column is associated to an edge. If the network is undirected, the incidence matrix is defined as

$$K_{ij} = \begin{cases} 1 & \text{if the node } u_i \text{ is incident to the edge } e_j, \\ 0 & \text{otherwise.} \end{cases}$$

If the network is directed, the incidence matrix indicates the origin of the edges with the value -1:

$$K_{ij} = \begin{cases} -1 & \text{if the node } u_i \text{ is the origin of the edge } e_j, \\ 1 & \text{if the node } u_i \text{ is the destination the edge } e_j, \\ 0 & \text{otherwise.} \end{cases}$$

An undirected network may be considered as directed by transforming each undirected edge into two directed edges of opposite origins and destinations. The incidence matrix of a directed network may be decomposed into the difference of two binary matrices so that $K = K_{\text{in}} - K_{\text{out}}$. This decomposition provides a relationship between the incidence matrix of a graph, its adjacency matrix A and the adjacency matrix G of its associated line graph through

$$A = K_{\text{out}} K_{\text{in}}^T \tag{1.1.1}$$

$$G = K_{\text{in}}^T K_{\text{out}}. \tag{1.1.2}$$

A third way of storing a network is the *edge list*, which lists the m couples of origin-destination associated to the edges (which extremity is considered as the origin being arbitrary in the case of undirected edges). An edge list is an appropriate way to encode a network on a computer but it is not convenient for mathematical development.

The degree d_i of a node u_i on an undirected network is the number of edges that are incident to it, and is given by

$$d_i = \sum_{j=1}^{n} A_{ij}. \tag{1.1.3}$$

If the network is directed, one similarly defines the in-degree and the out-degree of a node as the number of edges respectively entering and leaving the node.

The Laplacian matrix defined as $L = D - A$, where D is the diagonal matrix of the out-degrees of the nodes, allows one to describe some dynamics taking place on the network, as we will illustrate in Section 1.5.

The degree distribution $P(k)$ of a network, defined as the fraction of the nodes of degree k, is a standard tool for comparing networks and for detecting some organization in the network structure. The simplest model of random graph is the Erdős-Rényi model. This model assumes that each edge exists with the same probability p. The resulting degree distribution is a binomial with parameter p, which tends to a Poisson distribution of mean np when the number n of nodes is large. Therefore, a direct analysis of the organization of a network consists in comparing the degree distribution of a network to a Poisson distribution. A standard method for generating a random network is the configuration model that allows one to generate random networks from a given degree distribution.

The first moment $\langle k \rangle$ and the second moment $\langle k^2 \rangle$ of the degree distribution allow to detect correlations between the degree of adjacent nodes. For uncorrelated networks, the average degree of the neighbours of the nodes of degree k is given by $\frac{\langle k^2 \rangle}{\langle k \rangle}$ independently of k [Pastor-Satorras et al. 2001]. Therefore, if this average value varies with respect to k, it indicates a correlation between the degrees. The network is called assortative if this average value increases with respect to the degree and disassortative if it decreases.

A relevant question in the context of social networks is the identification of the nodes that have the greatest structural position in the network, relying on the notion of centrality of a node. Many different definitions of centrality have been developed depending on various indicators of influence. Each method assigns a score to each node and ranks them accordingly.

The simplest measure is the degree centrality, where the importance of a node is proportional to its degree. The closeness centrality [Noh and Rieger 2004] measures the mean shortest distance from a node to the others whereas the betweenness centrality [Freeman 1977] measures how often a node belongs to the shortest path between any pair of nodes. The eigenvector centrality [Ruhnau 2000] assumes that the importance of a node is proportional to the sum of the scores of its neighbours and is given by the normalized eigenvector associated to the largest eigenvalue of the adjacency matrix of the network [Ruhnau 2000]. One successful extension of the eigenvector centrality for directed network, is PageRank [Brin and Page 1998], used at the core of the Google search engine, which takes into account the in-going edges associated to a node and their weights.

Other measures of centrality rely on the concepts of k-core and k-shell decomposition [Seidman 1983]. The k-core is a maximal connected subgraph such that the degree of each node is larger than k. The nodes that are in the k-core but not in the $(k+1)$ -core form the k-shell. Therefore, these metrics favour the nodes that belong to a dense subgraph of the network.

Based on the problem one wants to solve, the appropriate centrality to use varies. For instance, let us consider a network where the nodes are the neighbourhoods of a city and the edges represent the fast connections between them. Then, the closeness centrality may be used to determine the ideal position of a firehouse, the betweenness centrality where to put an advertisement and the eigenvalue centrality where to book a touristic accommodation, as the latter would be closer to important neighbourhoods.

1.2 Complex networks

Empirical studies of real-world networks have shown that they are complex networks with non-trivial structural properties that differ from simple Erdős-Rényi networks, where each pair of nodes has the same probabilities of being connected. For instance, the degree distribution of the nodes deviates from the Poissonnian distribution of an Erdős-Rényi network, as it is typically heavy-tailed for scale-free networks [Barabási and Albert 1999], reflecting the presence of a small number of nodes which have a much larger degree than the others, often called hubs.

Moreover, the different connections between the nodes tend to correlate, as some groups of nodes are connected to each other more often that one might expect by chance, forming communities, or clusters, that possibly evolve and overlap [Girvan and Newman 2002; Fortunato 2010]. When several types of connections exist, the network may also exhibit a layered organization [Domenico et al. 2016].

Another typical characteristic of complex networks is the small-world effect, which reflects the fact that the shortest path between any pair of nodes is small compared to what one may expect [Watts and Strogatz 1998].

The availability of more detailed data allows one to take into account the temporal aspects of the connections. Networks may evolve and change over time, either because nodes or links are created and destroyed over time, or because they switch from active to inactive state. Temporal networks model networks where edges are activated from time to time. The time between two successive activations of an edge or a node is given by the inter-event distributions, which are empirically heavy-tailed [Barabási 2010].

1.3 Stochastic processess

Modeling the interactions between the different agents of a system allows a better understanding of the dynamical processes that involve these agents through the studied connections. The propagation of an entity *on* the network is indeed influenced by the structure of the network, but also by the interactions with the underlying dynamics *of* the network itself.

In order to take into account such temporal patterns of the networks, one associates a stochastic process to the nodes or to the edges. When the successive (positive) associated inter-event times are independent and identically distributed (i.i.d),

the corresponding process is called a *renewal* process. The renewal process is called a *Poisson process* when its associated inter-event time distribution is exponential, and a *semi-Markov process* when the inter-event times are not exponentially distributed but still i.i.d. A process is Markovian when its future state only depends on its current state regardless of its history, that is when the process is memoryless. As a consequence, the dynamics are Markovian only when the inter-events times are generated by a Poisson process. The i.i.d assumptions allow one to simplify many analytical derivations.

As an illustration, let us compute the probability density $P(k,t)$ to observe exactly k events in an interval of length t just after one event occurred at time $t = 0$. The time for the next event is given by the inter-event time distribution $\psi(t)$. Because of the independence, the time for k events to occur is given by the k convolutions of $\psi(t)$ with itself, denoted $\psi^{*k}(t)$. Because we consider positive random variables, the convolution product $*$ between two distributions f and g is given by

$$\begin{aligned}(f * g)(t) &= \int_{-\infty}^{+\infty} f(t')g(t-t')\,\mathrm{d}t' \\ &= \int_0^t f(t')g(t-t')\,\mathrm{d}t' \end{aligned} \tag{1.3.1}$$

The probability density $P(k,t)$ is given by the probability that the k^{th} event occurred at a time $t' < t$ and that no new event occurs in the interval $[t',t]$. Integrating over all the possible times t' leads to

$$P(k,t) = \int_0^t \psi^{*k}(t') \left(1 - \int_0^{t-t'} \psi(\tau) d\tau\right) \mathrm{d}t' \tag{1.3.2}$$

A standard way of dealing with expressions involving convolution products is to switch to the associated Laplace domain, as will be done in Chapter 5. The Laplace transform $\tilde{f}(s)$ of the function $f(t)$ is defined as

$$\tilde{f}(s) = \int_{0^-}^{+\infty} e^{-st} f(t) dt. \tag{1.3.3}$$

In the Laplace domain, a convolution product corresponds to a simple product through

$$\widetilde{(f * g)}(s) = \tilde{f}(s)\,\tilde{g}(s). \tag{1.3.4}$$

The Laplace transform $\tilde{g}$(s) of an integral $g(t) = \int_0^t \psi(\tau) d\tau$ is given by

$$\tilde{g}(s) = \frac{\tilde{\psi}(s)}{s}. \tag{1.3.5}$$

Therefore, the equation (1.3.2) simplifies in the Laplace domain to

$$\tilde{P}(k,s) = \tilde{\psi}^k(s)\,\frac{1-\tilde{\psi}(s)}{s}. \tag{1.3.6}$$

As we have illustrated, the i.i.d assumptions are convenient as they allow mathematical simplifications. However, these assumptions on the renewal processes do not hold in many human-related systems. For instance, the human activity varies according to daily or weekly cycles. Moreover, one interaction may trigger new ones, as a message may create a discussion between users, resulting in several correlated interactions. Poisson and semi-Markov processes are often used as baseline models which are compared to empirical data or more complex models. The study of temporal models often aims at quantifying the deviations from these simple models in order to understand the impact of a specific additional feature from the Markovian case. In this thesis, we will focus on diffusive and spreading dynamics on temporal networks when the dynamics in play are non-Markovian.

1.4 Burstiness

In many systems, agents exhibit bursty behaviour, or burstiness, which corresponds to the alternation of long periods of low activity and short periods of high activity [Karsai et al. 2017]. Such behaviour reflects non-Markovian dynamics, as the corresponding inter-event time distributions differ from an exponential distribution. Empirically, the corresponding distributions have been shown to be heavy-tailed and look like power-law distributions, in the sense that their complementary cumulative distributions functions are linear on a wide range when displayed in logarithmic axes. A more formal method for testing power-law distributions has been developed in [Clauset et al. 2009]. Bursty activities have been observed in a variety of natural phenomena, such as earthquakes [Bak et al. 2002], solar activity [Wheatland et al. 1998] or neuronal firing [Kemuriyama et al. 2010]. The availability and study of large-scale data has also unraveled that such bursty behaviour are also inherent to many human-related activities, such as (e)mail correspondence [Barabási 2005; Oliveira and Barabási 2005], mobile phone interactions through text messages or phone calls [Karsai et al. 2012], job submissions to clusters [Kleban and Clearwater 2003], online forums activity [Rocha et al. 2010] or taxis' mobility [Jiang et al. 2009].

The burstiness of a renewal process is measured using the mean $\langle \tau \rangle$ and the second moment $\langle \tau^2 \rangle$ or the variance σ^2 of the distribution of its inter-event times. Using the coefficient of variation of a distribution $CV = \frac{\sigma}{\langle \tau \rangle}$, the burstiness parameter $\mathbb{B}$ is defined as [Goh and Barabási 2008]:

$$\mathbb{B} = \frac{CV - 1}{CV + 1} \tag{1.4.1}$$

$$= \frac{\sigma - \langle \tau \rangle}{\sigma + \langle \tau \rangle}. \tag{1.4.2}$$

The burstiness parameter is equal to -1 for regular processes (dirac distributions), 0 for Markov processes, and close to 1 for highly bursty processes.

Other works have derived analytical expressions where another measure β of burstiness appears, also based on CV [Delvenne et al. 2015]:

$$\beta = \frac{CV^2 - 1}{2} \tag{1.4.3}$$

$$= \frac{\sigma^2 - \langle\tau\rangle^2}{2\langle\tau\rangle^2} \tag{1.4.4}$$

$$= \frac{\langle\tau^2\rangle}{2\langle\tau\rangle^2}. \tag{1.4.5}$$

Similarly, the measure β takes a value of $-\frac{1}{2}$ for regular processes, 0 for Markov processes, and large (unbounded) values for highly bursty processes.

Several models have been developed to explain the emergence of these bursty behaviour, and three main families of models emerge.

The initial model of burstiness was proposed in [Barabási 2005] and consists in a priority queuing model. The model considers a rational agent receiving tasks of different random importance and execution times. Then, the agent rationally executes the tasks following the order based on their priority. The distribution of the time between the insertion of a task in a list and its execution may be shown to be heavy-tailed, depending on the tuning of the parameters of the model, and the choice of the underlying arrival and execution time distributions.

Another family of models [Malmgren et al. 2008] explaining burstiness is based on the principle that human interactions depend on external factors with distinct scales, such as circadian or weekly patterns. This approach assumes that the activity alternates between Markovian and Non-Markovian processes.

A third family of models relies on the idea that the occurrence of an event acts as a trigger for the emergence of other events, and therefore that consecutive events are correlated. Such correlations may be induced using memory functions [Vazquez 2007], self-exciting [Masuda et al. 2013] or reinforcement processes [Wang et al. 2014]. One popular family of self-exciting process is the Hawkes process, initially introduced to describe earthquakes dynamics [Hawkes 1971], which assumes that the activity rate $\lambda(t)$ varies over time as

$$\lambda(t) = \lambda_0 + \sum_{t_i<t} \Phi(t - ti), \tag{1.4.6}$$

where $\Phi(t)$ is the memory Kernel, which allows to take into account the previous activities that occurred at times t_i.

1.5 Random Walks on temporal networks

The first part of this monograph deals with the concept of Random Walks on networks. The study of Random Walks has a long tradition in network science [Masuda et al. 2017]. Random Walks are at the heart of many algorithms to uncover central nodes [Brin and Page 1998; Gleich 2015] or communities of densely connected nodes [Rosvall and Bergstrom 2008; Delvenne et al. 2010], and they often serve as a first model to understand how the topology of a network, e.g. its degree distribution, affects diffusive processes [Chung 1996]. Random Walks have also been studied mathematically and numerically when the underlying topology is a network enriched with additional features. An important family of models in which we are particularly interested consists in temporal networks [Holme and Saramäki 2012; Masuda and Lambiotte 2016]. In this case, Random Walks are affected by the interplay between the network topology and the statistical properties of the node or edge dynamics [Starnini et al. 2012; Hoffmann et al. 2013; Delvenne et al. 2015]. Empirical observations have shown that the temporal processes associated to human-related networks strongly deviate from a Poisson process, due to their non-stationarity [Jo et al. 2012], to the correlations between the activation times of the network entities [Lambiotte et al. 2014; Karsai et al. 2012; Scholtes et al. 2014] or to the heavy-tailed distributions of the inter-event times of activation [Barabási 2010], to name a few reasons.

A Random Walk on a network is a stochastic process in which a walker is assumed to take a journey on a static underlying structure represented by a network [Newman 2010] with adjacency matrix A. Typically, the walker sits on a node until the decision to jump to a neighbouring node is taken according to some specific rules. There are two types of distinction among the models of Random Walks, depending on whether the renewal processes are associated to the nodes or the edges from on the one side, and whether the triggering of the renewal process are conditioned on the presence of the walker on the other side.

1.5.1 Active node-centric Random Walk

The active node-centric Random Walk is defined as follows. We assume that when the random walker arrives on a node i, it sits on it for a random duration t, drawn from a continuous waiting-time probability density function $\psi(t)$. Then, the walker jumps from i to one of its neighbouring nodes with equal probability. The waiting times between the different jumps being independent, the events are generated by a renewal process. For the sake of simplicity, we will assume the waiting time distributions to be the same on each node. The walk is called node-centric [Hoffmann et al. 2013; Speidel et al. 2015] because the renewal process is associated to the nodes, and active as the arrival of the walker on a node triggers the renewal process. A practical example may consist in a simplified model of population migration where each citizen is a random walker and may stay for a certain period in an area until deciding to move to a neighbouring place, according to some probability encoded through the weights of the adjacency matrix. Note that if the waiting-time distribution is discrete, the process

is in essence equivalent to a Markov chain where the states correspond to the nodes [Aldous and Fill 2002; Blum et al. 2016].

When the waiting times follow an exponential distribution ($\psi(t) = \lambda e^{-\lambda t}$), the process is Poisson and the diffusion is ruled by a linear differential equation where the probability density $\mathbf{n}(t)$ of the random walker being on the nodes evolves according to the flow induced by neighbouring nodes. From the transition matrix $T = D^{-1}A$, where D is the diagonal matrix of the out-degrees of the nodes, one defines the normalized Laplacian matrix $\tilde{L} = I - T$. Then, the dynamics is governed by the standard master equation

$$\dot{\mathbf{n}}(t) \quad = \quad -\lambda \tilde{L}\mathbf{n}(t). \tag{1.5.1}$$

The relaxation time, that indicates the rate of convergence to the stationary state, is determined by the spectral properties of the normalized Laplacian $\tilde{L}$, which is a semi-definite positive matrix.

However, in a large number of systems as diverse as mobile phone communication, email checking and brain activity, events taking place on the nodes exhibit complex temporal patterns and their dynamics tends to deviate strongly from a Markovian process [Liu et al. 2013; Holme and Saramäki 2012; Karsai et al. 2017]. In this case the general master equation is better expressed in the Laplace domain. Independently of the times of the jumps, the probability vector $\mathbf{p}(k)$ of the position of the walker after k jumps is given by

$$\mathbf{p}(k) \quad = \quad \mathbf{n}(0)T^k. \tag{1.5.2}$$

Then, the continuous time density $n(t)$ of the walker at time t may be computed by considering all the trajectories of different lengths and the probability $P(k,t)$ to perform exactly k jumps in a duration t

$$\mathbf{n}(t) \quad = \quad \sum_{k=0}^{\infty} \mathbf{p}(k)\, P(k,t). \tag{1.5.3}$$

The general master equation may be expressed in the Laplace domain as

$$\tilde{\mathbf{n}}(s) \quad = \quad \sum_{k=0}^{\infty} \mathbf{p}(k)\, \tilde{P}(k,s) \tag{1.5.4}$$

$$= \quad \frac{1-\tilde{\psi}(s)}{s} \sum_{k=0}^{\infty} \mathbf{p}(k)\, \tilde{\psi}^k(s) \tag{1.5.5}$$

$$= \quad \frac{1-\psi(s)}{s} \mathbf{n}(0)[I - T\tilde{\psi}(s)]^{-1}. \tag{1.5.6}$$

where the second equality is obtained using equation (1.3.6) and the third using equation (1.5.2) and the Neumann's lemma applied to the stochastic matrix T.

The estimation of the impact of the deviation from a Poisson process on the relaxation of the different eigenmodes of T allows to determine how the shape of the waiting-time distribution affects the speed of the diffusion [Delvenne et al. 2015].

From the transition matrix T, it is possible to construct an orthonormal basis of eigenvectors. Denoting $\mathbf{v}_{l,R}$ (resp. $\mathbf{v}_{l,L}$) the right (resp. left) normalized eigenvector associated to the eigenvalue λ_l, the k^{th} power of the transition matrix T is given by

$$T^k = \sum_{l=1}^{n} \lambda_l^k \mathbf{v}_{l,R} \mathbf{v}_{l,L}. \tag{1.5.7}$$

Plugging (1.5.7) into (1.5.2), it is thus possible to express the density of the walker after k jumps depending on the eigenvectors and eigenvalues of T as

$$\mathbf{p}(k) = \sum_{l=1}^{n} \lambda_l^k c_l \mathbf{v}_{l,L}, \tag{1.5.8}$$

where $c_l = \left\langle \mathbf{n}(0), \mathbf{v}_{l,R} \right\rangle$.

Plugging this new expression of $\mathbf{p}(k)$ into equation (1.5.5) leads to

$$\tilde{\mathbf{n}}(s) = \frac{1-\tilde{\psi}(s)}{s} \sum_{l=1}^{n} c_l \frac{1}{1-\lambda_l\, \tilde{\psi}(s)} \mathbf{v}_{l,L}, \tag{1.5.9}$$

Exploiting the orthonormal basis of eigenvectors, one may multiply both members of the equation (1.5.9) by $\mathbf{v}_{l,R}$ in order to obtain the evolution of the eigenmode $\tilde{a}_l(s)$ associated to λ_l in the Laplace domain as

$$\tilde{a}(s) = \frac{1-\tilde{\psi}(s)}{s\,(1-\lambda_l\, \tilde{\psi}(s))} c_l \tag{1.5.10}$$

$$\approx c_l \frac{\langle t \rangle}{1-\lambda_l} \left(\frac{1 - \frac{\langle t^2 \rangle}{2\langle t \rangle} s}{1 + \frac{\lambda_l}{1-\lambda_l} s \langle t \rangle - \frac{\lambda_l}{1-\lambda_l} \frac{s^2}{2} \langle t^2 \rangle} \right) \tag{1.5.11}$$

$$\approx c_l \frac{\langle t \rangle}{1-\lambda_l} \left(1 - s\left(\frac{\langle t^2 \rangle}{2\langle t \rangle} + \frac{\lambda_l}{1-\lambda_l} \langle t \rangle \right) \right), \tag{1.5.12}$$

where the equation (1.5.11) is obtained using the second-order small s expansion of $\psi(s)$, which is given by

$$\tilde{\psi}(s) = 1 - s\langle t \rangle + \frac{s^2}{2} \langle t^2 \rangle + \mathcal{O}(s^3), \tag{1.5.13}$$

and the equation (1.5.12) is obtained using the following first order approximation around $s = 0$

$$\frac{1-as}{1+bs-cs^2} = 1 - s(a+b) + \mathcal{O}(s^2). \tag{1.5.14}$$

The relaxation of an eigenmode associated to $\lambda < 1$ is therefore governed by three terms, appearing in the characteristic time t_{ch} given by

$$t_{ch} = \langle t \rangle \left(\frac{1}{1-\lambda} + \beta\right), \tag{1.5.15}$$

where $\beta = \frac{\langle t^2 \rangle}{2\langle t \rangle^2}$ is a measure of the burstiness of the process introduced in section 1.4, and $1-\lambda$ is the spectral gap of the transition matrix T.

This result allows one to determine which of the structure or the temporal patterns of the network impacts the most the speed of diffusion. A small spectral gap $1-\lambda$ implies the existence of bottlenecks or well-defined communities in the underlying network [Lovász 1993]. In this case, the topology governs the speed of relaxation. When the spectral gap is larger, or when the burstiness of the process is very large, the speed of relaxation is governed by the properties of $\psi(t)$, in particular through its tail and its variance. This result shows that burstiness generally tends to slow down the diffusion on networks.

1.5.2 Active edge-centric Random Walk

The active edge-centric Random Walk differs from the active node-centric Random Walk in the sense that the renewal processes are associated to the edges instead of the nodes. In this standard framework of temporal networks, which is a natural model for contacts in social networks, we assume that when the random walker arrives on a node i, it triggers on each incident outgoing edge an associated random waiting time for the so-called 'activation' of the edge from the associated inter-activation distribution, also called inter-event distribution. The random walker jumps through the first edge that reaches activation. As we consider continuous inter-activation, the probability that two edges are activated simultaneously is zero (almost surely) and the walker never has to choose between multiple available edges. After a jump, the process restarts on the new node reached by the random walker. Here, activation is considered as an event of infinitesimal duration, allowing the passage of the random walker. The time at which each edge ij reaches activation is independently drawn from its own inter-event time distribution $f_{ij}(t)$. For the sake of simplicity, we will assume the activation time distribution $f(t)$ to be the same on each edge.

Such model corresponds to a situation where the clock is associated to the moving entity, whose arrival triggers a waiting-time that will differ on each edge. Mathematically, this model may be seen as an extension of the active node-centric RW. Back to the migration example from the node-centric paradigm, one may consider here that moving opportunities appear asynchronously for the neighbouring places.

Again, when the process is Poisson, the spreading is described by a linear differential equation, corresponding to the heat equation for a continuous medium to a network [Lawler 2010]. The master equation for the dynamics involves this time the Laplacian matrix $L = D - A$ through

$$\dot{\mathbf{n}}(t) \quad = \quad -L\mathbf{n}(t). \tag{1.5.16}$$

Similarly to the normalized Laplacian $\tilde{L}$, the Laplacian L is a semi-definite positive matrix.

When the process is not Markovian, the process may still be seen as the result of a competition between the different independent edges, from the walkers' outlook. The walker will wait until the fastest edge is activated, and the waiting-time distribution $\psi_i(t)$ of leaving the node i of degree d_i is given by the distribution of the minimum $X_{(1)}$ of d_i i.i.d random variables following the same distribution $f(t)$. The repartition function of $X_{(1)}$ is given by

$$P(X_{(1)} \leq t) = 1 - (1 - P(X \leq t))^{d_i}, \tag{1.5.17}$$

where X is a random variable with distribution $f(t)$.

Derivating the equation (1.5.17) provides the waiting-time distribution $\psi_i(t)$ as

$$\psi_i(t) \quad = \quad d_i P(X \leq t)(1 - P(X \leq t))^{d_i - 1} \tag{1.5.18}$$

$$\quad = \quad d_i f(t) \left[\int_t^{+\infty} f(\tau)\, \mathrm{d}\tau \right]^{d_i - 1}. \tag{1.5.19}$$

Thus, when the edges are statistically equivalent, the active edge-centric Random Walk may be associated to an active node-centric Random Walk with heterogeneous waiting-time distributions.

1.5.3 Passive edge-centric Random Walk and the bus paradox

The previous models of active Random Walks assume that the arrival of the walker on a node triggers the renewal process associated to the nodes or the edges. However, the activity of the edges happens to be independent from the walker in many real-life situations of contact-based networks [Holme and Saramäki 2012; Karsai et al. 2017]. In other words, the dynamics *of* the network is independent from the entity that spreads with its own dynamics *on* the network. The Random Walk is called passive when contacts are taking place on edges regardless of the presence of a random walker. The random walker waiting on a node jumps through the first available edge incident to the node. Similarly to the active edge-centric Random Walk, edges are activated for an infinitesimal duration and the time between two consecutive activations of an edge is governed by a renewal process, with i.i.d. inter-activation times distributed according to a probability density function. The activation processes on different edges are independent and we consider again the same inter-activation distributions $f(t)$ for every

edge; nonetheless the forthcoming analytical derivations hold for distinct edge activation distributions as well. Such a model may be used in order to model information diffusion on a communication network, where contact events are independent from the message passing through the network.

In order to describe the process, it is crucial to estimate the waiting time distribution $g(t)$ of the walker on each edge, i.e. the time that the random walker arriving on a node has to wait before a given edge is activated. The relation between the inter-activation distribution $f(t)$ on a given edge and the waiting-time distribution $g(t)$ is given by the so-called bus paradox. This classical result in stochastic process and queuing theory is also named waiting-time paradox or inspection paradox [Feller 1971; Çinlar 1969]. The apparent paradox arises when one computes the mean of the waiting time distribution $g(t)$.

The waiting time distribution $g(t)$ is derived as follows (see illustration on Figure 1.1).

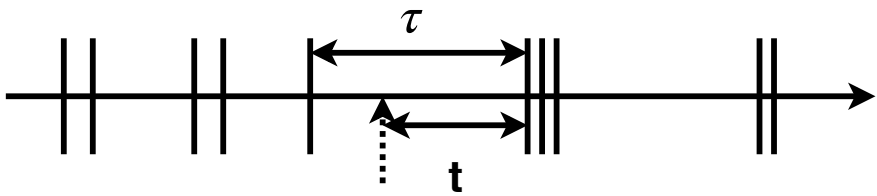

Figure 1.1 – The distribution of the waiting-time t when randomly arriving between two activations may be derived from the distribution of the inter-activation times τ.

Under the assumption that the arrival of the walker on the node is independent from the edge activation, the probability density to fall into an inter-event interval of length τ is proportional to τ and to the probability function $f(\tau)$ that such an event of length τ occurs, normalized over all the possible inter-event length τ'. With no additional information, the probability to fall anywhere on the interval is $\frac{1}{\tau}$. The resulting distribution $g(t)$ is obtained by integrating over all the intervals of length $\tau > t$, which yields to

$$g(t) = \int_t^{+\infty} \frac{\tau f(\tau)}{\int_0^{+\infty} \tau' f(\tau')\,\mathrm{d}\tau'} \frac{1}{\tau}\,\mathrm{d}\tau \tag{1.5.20}$$

$$= \frac{1}{\langle\tau\rangle} \int_t^{+\infty} f(\tau)\,\mathrm{d}\tau, \tag{1.5.21}$$

where $\langle\tau\rangle \equiv \int_0^{+\infty} \tau f(\tau)\,\mathrm{d}\tau$ is the mean inter-activation time.

Notably, when the distribution of f is exponential, g is equal to f and the resulting dynamics of the walk is exactly the same than an active edge-centric Random Walk.

The mean waiting-time of $g(t)$ is computed by permuting the two integrals:

$$\int_0^{+\infty} t g(t)\,\mathrm{d}t = \int_0^{+\infty} t \left[\frac{1}{\langle\tau\rangle}\int_t^{+\infty} f(\tau)\,\mathrm{d}\tau\right]\mathrm{d}t \tag{1.5.22}$$

$$= \frac{1}{\langle\tau\rangle}\int_0^{+\infty} f(\tau)\left[\int_0^{+\infty} t\,\mathrm{d}t\right]\mathrm{d}\tau \tag{1.5.23}$$

$$= \frac{\langle\tau^2\rangle}{2\langle\tau\rangle}, \tag{1.5.24}$$

$$= \frac{\langle\tau\rangle}{2} + \frac{\sigma^2}{2\langle\tau\rangle}, \tag{1.5.25}$$

where $\langle\tau^2\rangle$ is the second moment of the inter-event distribution f, and σ^2 its variance.

Thus the paradox lies in the fact that the mean waiting-time is larger than the half of the mean inter-event time as soon as the process is not periodic. In situations where the inter-event distribution presents a heavy tail, the mean waiting-time can even become arbitrarily large as compared to the average inter-event time. This result, which seems paradoxical at first glance, originates from the fact that the inter-contact interval duration are sampled proportionally to their duration.

The situation is also called the bus paradox [Avineri 2004] because of the following analogy: consider a person arriving randomly at a bus stop and waiting for the next bus to show. Then, the waiting-time at the bus stop will be on average larger than the half of the scheduled time between two buses, and particularly large when the bus service is bursty.

The obtained waiting-time distribution $g(t)$ may be used to reduce the passive edge-centric Random Walk to an active edge-centric Random Walk, however this generalization only holds for directed graphs with no short cycles. Indeed, the available master equations are only approximate as one looses the independence assumption between the renewal processes associated to each edge as soon as the walker has the possibility to jump to an already visited node, as memory will emerge in the path of the walker. The fact to jump back to the last visited node is called backtracking. We will study in chapter 2 the emergence of non-Markovian trajectories under such backtracking bias inherent to the non-Markovian passive edge-centric Random Walk.

1.5.4 Extensions of edge/node-centric Random Walk

Of course, dynamics on node and edges are not mutually exclusive, and one may allow the co-existence of waiting-times associated to both the nodes and the edges. One may consider for instance the diffusion of a virus on a social network when the virus first needs to incubate before the newly infected individual becomes contagious, or about a tourist visiting different cities through public transportation and spending some time to explore the city.

Assuming homogeneous waiting-time distribution $f(t)$ and inter-activation distributions $\psi(t)$, the combination of an edge-centric and an active node-centric model may be equivalent to an active node-centric Random Walk.

If the edge-wise component is active, the associated waiting-time $\psi_i^a(t)$ on node i is given by

$$\psi_i^a(t) \quad = \quad (\psi * f_{(1),i})(t), \tag{1.5.26}$$

where $*$ stands for the convolution product, and $f_{(1),i}$ denotes the distribution of the minimum of $d(i)$ random variables with distribution f.

In the situation where the edge-wise component is passive, the associated waiting-time $\psi_i^p(t)$ on node i given by

$$\psi_i^p(t) \quad = \quad (\psi * g_{(1),i})(t), \tag{1.5.27}$$

where $g_{(1),i}$ denotes the distribution of the minimum of $d(i)$ random variables with distribution g defined following equation (1.5.21). Moreover, if the distributions associated to the edges are not equal, the reduction does not hold since it implicitly assumes that edges are statistically equivalent.

However, it is important to keep in mind that the latter reduction only holds for directed graphs with no short cycles, again because of the emergence of memory in the walk when short cycles are in play.

Finally, all these models assume infinitesimal edges'activation, which is an assumption that holds when the duration of the contacts is negligible with respect to the studied dynamics. For instance, the time required to read a message is very small compared to the time between distinct messages, or the duration of a phone call is very small compared to the length of a longitudinal study of call data record [Hoffmann et al. 2013; Karsai et al. 2017]. However, such assumption does not hold in situations where the different timescales of and on the network are of the same order. Taking into account the co-existence of both dynamics on a continuous-time Random Walk will be the subject of Chapter 4.

1.5.5 Random Walks on multilayer networks

Another important family of models of networks is made of multilayer networks [Kivelä et al. 2014; Gómez et al. 2013; Domenico et al. 2016; Aleta and Moreno 2019], where different types of connections exist between the nodes, as in social networks [Szell et al. 2010; Magnani et al. 2013] or transportation networks [Cardillo et al. 2013; Gallotti and Barthelemy 2015] for instance. These networks have a layered organization and are usually represented as tensors or by means of a so-called supra-adjacency matrix. Random Walks have also been studied in this context, to uncover how the presence of multiple layers affects diffusion [Domenico et al.; Boccaletti et al. 2014] or to define generalized versions of Pagerank [De Domenico et al. 2015]. Multiplex networks are multilayer networks where the only connections between nodes from different layers involve two duplicates of the same node.

Let us consider the Markovian edge-centric Random Walk on a network composed of two layers. Denoting L_i the associated Laplacian of each layer i and D_X the interlayer diffusion constant, the master equation of the walk is similar to equation 1.5.16:

$$\dot{\mathbf{n}}(t) \quad = \quad -\mathbf{n}(t)\mathscr{L}, \tag{1.5.28}$$

where $\mathscr{L}$ is the supra-Laplacian of the graph defined as

$$\mathscr{L} \quad = \quad \begin{bmatrix} L_1 + D_X I & -D_X I \\ -D_X I & L_2 + D_X I \end{bmatrix} \tag{1.5.29}$$

In this thesis, we will only consider noninterconnected (or edge-colored) multiplex networks [Domenico et al.], a particular case of multilayer network, also called multigraphs [Newman 2010]. In such networks, the required time to switch between layers is negligible compared to the one required to move between neighbours. Thus, the duplicate of the nodes of the different layers may be merged and only the edges are associated to different layers (one layer corresponding to one color). This is the case for instance when one considers social interactions between users through different online social networks such as Facebook or Twitter, as the switching times between the channels are very low.

The coverage $\rho(\nu)$ of a Random Walk on a multiplex network is defined as the percentage of distinct nodes visited by the walker after ν jumps. The study of the coverage allows to compare Random Walks driven by different dynamics in terms of trajectory, regardless of the speed of the process. On an undirected interconnected multiplex network of size N with α layers, the coverage may be approximated as follows [Domenico et al. 2014]. Similarly to the continuous time framework, let us denote $\mathbf{n}(\nu)$ the probability vector $(1 \times \alpha\nu)$ of the walker to be at a given node after ν jumps. The probability $p_i(\nu)$ for the walker to be at node i after ν steps irrespective of the layer is given by

$$p_i(\nu) \quad = \quad \sum_{j=1}^{\alpha} n_{i+(j-1)N}(\nu) \tag{1.5.30}$$

$$= \mathbf{n}(\nu)\mathbf{E_i}^T \quad (1.5.31)$$
$$= \mathbf{n}(0)T^{\nu}\mathbf{E_i}^T, \quad (1.5.32)$$

where T is the block-diagonal transition matrix ($\alpha\nu \times \alpha\nu$), $\mathbf{n}(0)$ the vector of initial conditions and $\mathbf{E_i}$ a vector ($1 \times \alpha\nu$) corresponding to α concatenation of the standard canonical vector $\mathbf{e_i}$.

The probability $h_i(\nu)$ that the walker has not visited the node i after ν steps obeys the recursive relation

$$h_i(\nu+1) = h_i(\nu)(1-p_i(\nu+1)) \quad (1.5.33)$$
$$= h_i(\nu)(1-\mathbf{n}(0)T^{\nu+1}\mathbf{E_i}^T), \quad (1.5.34)$$

which leads to

$$\dot{h}_i(\nu) = -h_i(\nu)\mathbf{n}(0)T^{\nu}\mathbf{E_i}^T. \quad (1.5.35)$$

The solution of equation (1.5.35) is given by

$$h_i(\nu) = h_i(0)\exp(-\mathbf{n}(0)\mathbb{T}_{\nu}\mathbf{E_i}^T), \quad (1.5.36)$$

where $\mathbb{T}_{\nu} = \sum_{k'=0}^{\nu} T^{k'+1}$ is a matrix that takes into account all the paths of length ranging from 0 to $k+1$.

Finally, the coverage $\rho(\nu)$ of the walker after ν steps is approximated after averaging $h_i(\nu)$ over all the nodes i, and over all the possible initial conditions $\mathbf{F_j} = (\mathbf{e_j}, 0, \ldots, 0)$ for $j = 1, \ldots, n$, which leads to

$$\rho(\nu) = 1 - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} h_i(\nu|\mathbf{n}(0) = (\mathbf{e_j}, 0, \ldots, 0)) \quad (1.5.37)$$
$$= 1 - \frac{1}{N^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{N} \exp(-\mathbf{F_j}\mathbb{T}_{\nu}\mathbf{E_i}^T). \quad (1.5.38)$$

In particular, we will study in chapter 3 the edge-centric Random Walk on a multiplex network when the distributions of the inter-activation times associated to the edges are identical for the edges belonging to the same layer, but which may vary across different layers. We will show how the inter-layer activity heterogeneity may induce memory in the trajectory of the random walker and impact the network coverage of the walker.

Table 1.1 summarizes the distinctions between the active and passive Random Walks in terms of the waiting-time distribution associated to their renewal process and in terms of the possible emergence of different effects that we will study in the chapters 2, 3 and 4. Table 1.2 specifies the conditions under which a passive (resp. edge-centric) RW may be reduced to an active (resp.) node-centric RW, an active node-centric RW being the simplest RW model.

	Active RW	Passive RW
Associated Renewal Process	Triggered by the random walker	Independent of the random walker
Waiting-time distribution on an edge ij (edge-centric RW)	$f_{ij}(\tau)$	$\frac{1}{\langle\tau\rangle}\int_t^{+\infty} f_{ij}(\tau)\,\mathrm{d}\tau$
Waiting-time distribution at a node i (node-centric RW)	$\psi_i(x)$	$\frac{1}{\langle x\rangle}\int_r^{+\infty} \psi_i(x)\,\mathrm{d}x$
Backtracking Bias (if non-Markovian)	No	Yes (for undirected networks)
Memory induced by short-cycles	No	Yes
Rock-Paper-Scissors effect (on multiplex networks)	Yes	Yes (with possible opposite bias)

Table 1.1 – Main distinctions between active and passive RW.

Model Reduction of RW	Conditions
Passive $\rightarrow$ Active	Directed network without loops (DAG)
Edge-centric $\rightarrow$ Node-centric	Identical waiting-time distributions associated to the edges leaving the same node

Table 1.2 – Conditions for RW model reductions

1.6 Spreading models

The second part of this thesis deals with spreading models and how to leverage on the learnings from the previous study of the stochastic process in the context of temporal network in order to adapt one's spreading strategy. Epidemics processes are used as models to capture the dynamics of information or disease spreading. The main distinction between the diffusion of random walkers on a network and the spreading of information or disease over a network lies in the fact the the first process is conservative whereas the second assumes the replication of the spreading entity. As a consequence, spreading processes are immune to cycles trapping or backtracking bias.

1.6.1 Compartmental models

The simplest models of disease spreading are the SI, SIS and SIR compartmental models [Kermack and McKendrick 1927; Wearing et al. 2005] which assume that the individuals may belong to different compartments corresponding to their possibles states: susceptible (S), infected (I) or possibly recovered (R).

In the SI model, the individuals may switch from compartments according to the single following mechanism. When an infected individual contacts a susceptible one, the latter may get infected given some infection rate β:

$$S + I \quad \stackrel{\beta}{\rightarrow} \quad 2I \tag{1.6.1}$$

The SIS and SIR model incorporate a second mechanism in addition to (1.6.1): the infected individual may leave the infected state at a given recovery rate μ. The SIR model assumes a transition from the infected to the recovered state (1.6.2), meaning that no reinfection may occur, whereas the SIS model assumes a transition back from the infected to the susceptible state (1.6.3), allowing a reinfection of the individuals:

$$I \quad \stackrel{\mu}{\rightarrow} \quad R \quad \text{for the SIR model,} \tag{1.6.2}$$

$$I \quad \stackrel{\mu}{\rightarrow} \quad S \quad \text{for the SIS model.} \tag{1.6.3}$$

One natural goal in the study of epidemic models is to determine the epidemic threshold above which a pandemic will occur [Chakrabarti et al. 2008].

When the population is well-mixed, that is in the case of a complete network, the epidemic threshold is the same for the SIS and SIR models and given by $\frac{\beta}{\mu} = 1$.

For general networks, a linear stability analysis of the differential equations associated to the process allows one to show that the epidemic threshold is given by $\frac{\beta}{\mu} = \frac{1}{\lambda}$, where λ is the largest eigenvalue of the connectivity matrix associated to the network and the process [Boguñá and Pastor-Satorras 2002]. The connectivity matrix provides the number of available neighbours of degree j an infected node of degree j may infect, and thus depends on the degree distribution $P(k)$.

For uncorrelated networks, the connectivity matrix may be determined using the degree mean-field approach, which considers that the two nodes of equal degree are statistically equivalent. For the SIS model, each of the $jP(j)$ edges incident to a node of degree j is incident to another node proportionally to its degree i. One needs to normalize by the mean degree $\langle k \rangle$ because the sum of the i^{th} line corresponds to the total number of connections of a node of degree i, thus is equal to i. Therefore, the element C_{ij} of the connectivity matrix of an uncorrelated network [Newman 2010; Pastor-Satorras et al. 2015] is given by

$$C_{ij} \quad = \quad i\frac{jP(j)}{\langle k \rangle}. \tag{1.6.4}$$

The only non-zero eigenvalue of C is $\frac{\langle k^2 \rangle}{\langle k \rangle}$ and an associated eigenvector v is such that $v_i = i$ as shown below

$$(C\, v_i)_i = \sum_{j=1}^{n} i\frac{jP(j)}{\langle k\rangle} v_j \tag{1.6.5}$$

$$= \frac{i}{\langle k\rangle}\sum_{j=1}^{n} j^2 P(j) \tag{1.6.6}$$

$$= v_i \frac{\langle k^2\rangle}{\langle k\rangle}. \tag{1.6.7}$$

For the SIR model on an uncorrelated network, the connectivity matrix $\tilde{C}$ takes into account the fact that it is not possible to infect the node from which the infection arrived. Its element are given by

$$\tilde{C}_{ij} = i\frac{(j-1)P(j)}{\langle k\rangle}. \tag{1.6.8}$$

A similar development as in equation (1.6.7) shows that the only non-zero eigenvalue of $\tilde{C}$ is $\dfrac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}$.

In both the SIS and the SIR models, the epidemic threshold in terms of $\frac{\beta}{\mu}$ depends on a ratio between the first and the second moment of the degree distribution. As we already mentioned, the degree distribution of social networks is empirically heavy-tailed [Barabási and Albert 1999], which implies that $\langle k^2\rangle >> \langle k\rangle$. Therefore, the resulting epidemic threshold may be very small.

There exist many variations of the SI and SIR models, among which the most reputed are the SEIR (which incorporates an incubation period) and the SIRS (when the recovered individual may get infected after some immunity time). These extended models consider more compartments and allow to consider specific scenarios, such as at risk population or a time-limited disease. Other developments take into account non-Markovian interactions, such as correlations between successive interactions or non-exponential distributions of the recovery and transmission times. Compartmental models have also been used as baseline models in the context of information diffusion, where being infected corresponds to having reshared the content.

1.6.2 Models of cascade diffusion

In recent times, micro-blogging systems such as Twitter have become influential for spreading and sharing breaking news, personal updates and spontaneous ideas. For instance, Twitter has served as an effective medium for real-time posts about earthquakes, epidemic outbreaks or spreading awareness in many situations, like the Arab-Spring movement in 2011 [Wolfsfeld et al. 2013] or the U.S. presidential elections in 2016 [Bovet and Makse 2019]. Twitter provides retweeting facility, through which

a tweet simply gets relayed to all the followers of the retweeting user. The retweets from the original tweet form a cascade that spreads over the underlying social network. Specific models have been developed to predict the future size and form of the cascade tree. There are two types of models for cascade diffusion, based on features or on point-process methods.

Feature based machine learning methods rely on a list of potentially relevant features for each cascade, such as content, meta information about the users, structural and temporal features [Cheng et al. 2014]. By using learning algorithms or statistical methods on these features, cascades are classified and their future size is predicted. For instance, using various structural and temporal features of cascades, [Weng et al. 2014] showed that the initial diversity of a cascade across several network communities is a good predictor for future popularity, while [Yang and Counts 2010] studied the speed and scale of cascades, and [Pramanik et al. 2016] have investigated the specific role of mentions on Twitter. Studying a sample of large photo cascades on Facebook, [Cheng et al. 2016] showed that temporal and structural features allow to predict the future growth of a cascade and its potential resurgence.

Other works have specifically focused on the shape of the cascades. [Gómez et al. 2012] proposed a generating model for discussion threads based on preferential attachment, which they enriched by exploiting content quality. [Goel et al. 2015] have introduced the concept of structural virality that allows to formally discriminate between flat (star-like) and deeper cascades, the latter tending to be more associated to fake-news diffusion [Vosoughi et al. 2018].

However, such methods have important drawbacks: an extensive training is required, the performance highly depends on the quality of the features extraction [Suh et al. 2010] and the models typically rely on features that are time-consuming to extract [Bandari et al. 2012].

On the contrary, point-process methods, and in particular Hawkes processes, directly model the formation and diffusion process of the cascade. The goal is to calibrate a model of the time series associated to the cascade over a given period. These models rely on multiplicative effects, as new events tend to trigger new ones, reflecting the fact that a retweet may expose the content to a new population. The advantage of such methods is that they do not require feature engineering and may be used to predict the final outcome of tweets that are still spreading on the network.

In this vein, Gao et al. [Gao et al. 2015] introduced a deterministic time-dependent Poisson process model which was extended with the *SEISMIC* model (self-exciting Model of Information Cascades) [Zhao et al. 2015] method by taking into account the underlying structure of the network and by assuming that the activity rate $\lambda(t)$ depends on the initial post infectiousness, which stochastically changes over time. The *TiDeH* model (Time-Dependent Hawkes Process) [Kobayashi and Lambiotte 2016] outperformed *SEISMIC* by taking into account the circadian rhythms of online popularity and the aging of information, and also allowed to predict the full shape of the size evolution of the cascade, in contrast with the previous methods that only focused on predicting its final size.

1.6.3 Identification of influential users

A complementary approach to the study of cascades diffusion is the detection of influential users. Identification of such users has attracted widespread research interest over the years. Targeting these users allows one to facilitate the diffusion of information at a smaller cost, or to protect the population by vaccinating them in the context of disease spreading. Various centrality measures based on the underlying topology have been developed in order to detect influential users. For instance, [Huang et al. 2014] have identified important nodes based on their network roles, such as core or bridge, by combining multiple indicators with strong correlations while [Sheikhahmadi and Nematbakhsh 2017] have combined the degree, the core number of a node as well as the weighted diversity in the core number of friends to identify influential nodes. [Xia et al. 2016] suggested to exploit metadata to target users with a specific set of characteristics based on advertisers' preferences. Initially introduced by [Burt 1993], structural holes, acting as bridges connecting separated parts of a social network, have been investigated in several works and their detection techniques have been developed exploiting extra structural features [Ding et al. 2016; Lou and Tang 2013; He et al. 2016].

Another approach consists in exploiting direct influence measures in order to rank the users. For instance, [Malliaros et al. 2016] have proposed the K-truss decomposition of a graph based on triangle-based extension of k-core decomposition method [Carmi et al. 2007] to identify influential users while [Al-garadi et al. 2017] have proposed a link-weighting k-core decomposition method based on user interactions to identify influential spreaders. [Jianqiang et al. 2017] have proposed a measure of influence based on the Random Walk theory to identify the most influential users in a network by combining user influence based on the contribution of its tweets as well as its position in the network by combining degree, closeness and betweenness centrality measures. [Madotto and Liu 2016] have identified the super spreader nodes that maximize impact on other nodes by ranking the users through combining eight different centrality measures using a modified Borda count aggregation on a variety of real-world networks.

However, these methods only rely on the structural properties of the networks, neglecting the temporal aspects of the cascades spreading on it, and in particular overlooking the potential of inter-retweet intervals in a cascade that may provide important signatures to identify a better set of influential nodes in a network, when combined with structural information. In Chapter 5, we will analytically study how the inter-activation time distribution impacts the spread of the transmission of a disease on the SI model when several attempts are necessary in order to succeed the transmission from an infected individual to a susceptible one. Building on Twitter cascades datasets, we will further show in Chapter 6 that the time series of a retweet cascade may provide information on how it has spread on the underlying network. We will present in Chapter 7 how to exploit this latter findings in order to address, at a lower cost and with better efficiency than the standard methods, the specific goal of multi-seeds targeting in order to maximize the spread of a message diffusion, when one may choose the initial seeds.

1.7 Data collection on Twitter

The first major problem one faces in the study of empirical data is the collection of the data itself.

The collection of data on Twitter requires the use of the query-based Twitter application programming interface (API). Its access is facilitated using specified modules, such as *Tweepy* in Python [Roesslein 2019], or *rtweet* in R [Kearney 2019]. One id is assigned to each tweet and to each user. The Twitter API allows one to obtain tweets containing specific keywords, respecting specific geolocation based criteria or from a specific tweet id. It also provides profile information and ego social network from a user id. In order to use the Twitter API, one first needs to register the application on Twitter (through `dev.twitter.com`). Once registered, one receives an API key and an API secret, which are not linked to a user but allow one to authenticate and send requests to the Twitter API. Any tweet is by default public and may be collected. However, free account have several limitations. First, the search API of Twitter only allows to detect tweets emitted in the last 7 days. Nevertheless, it is possible to crawl an older tweet based on its tweet id. Second, if the request query is too large, for instance concerning all the tweets of a very popular hashtag or a large set of users, the set of collected data will be incomplete. Moreover the *Get* query calls are limited per time-window (15 calls in May 2019, see `https://developer.twitter.com/en/docs/basics/rate-limiting`), a time-window lasting 15 minutes. However, this rate-limit may be tackled by deploying distributed crawler, for instance via PlanetLab.

Many datasets are publicly available for instance through the following links[(1)]:
`http://dfreelon.org/2017/01/03/beyond-the-hashtags-twitter-data/`,
`https://www.docnow.io/catalog/`,
`https://data.world/datasets/twitter`.
Twitter's Developer Policy only allows one to share the ids of the tweets and of the users of a datasets, which means that one needs to crawl the data from the tweets and users ids. Such a limitation prevents deleted or newly protected tweets to be shared.

Finally, it is worth mentioning to the interested reader that, in the context of the collaboration that lead to the development of *SmartInf*, Ayan Bhowmick provided on Github a script in Python allowing to crawl tweets related to specific hashtags, which is available at `https://github.com/ayan-0305/SmartInf`.

[(1)] All the url links provided in this section have been verified in July 2019

Part II

Random Walks on temporal networks: emergence of memory

Chapter 2

Backtracking and Mixing Rate on uncorrelated temporal networks

This chapter presents the results of [Gueuning et al. 2017].

Abstract

We consider the passive edge-centric Random Walk as defined in section 1.5.3. Despite the simplicity of the model, we show how the random walker's trajectory is affected by its emerging memory. In particular, we quantify the walker's tendency to backtrack, as well as the resulting effect on the mixing rate of the process. As we show through empirical data, non-Markovian dynamics may significantly slow down diffusion due to the backtracking. Such effect is linked to the bus paradox but intrinsically differs from it. We conclude by discussing the implications of our work for the interpretation of results generated by null models of temporal networks.

2.1 Introduction

As we already mentioned, a central question in the study of diffusion on temporal network is the understanding of the mechanisms that either accelerate or slow down the diffusion, for instance through the characteristic time for the dynamics to converge to the equilibrium state. This question has been considered by means of numerical simulations, by simulating a diffusive process on empirical temporal network data [Starnini et al. 2012], and comparing its speed with the same process run on randomized null models [Karsai et al. 2011; Rocha et al. 2011]. A theoretical approach, which we also adopt here, consists in neglecting correlations between the activations of different edges, and modelling their dynamics as independent renewal processes [Hoffmann et al. 2013; Speidel et al. 2015], corresponding to the passive edge-centric Random Walk defined in section 1.5.3. In particular, we explore in detail the implications of an apparently paradoxical situation [Speidel et al. 2015; Saramäki and Holme 2015]: despite the fact that edges are independent processes, they cease to be independent along the path of a walker when the inter-event time distribution is non-exponential, which may lead to biases in its dynamics and non-Markovian trajectories.

In this chapter we illustrate and analyze this effect for a specific dependency pattern between successive jumps, namely the tendency for the random walker to backtrack, i.e. return to the previously visited node more than a purely Markovian walker would. Our contributions are twofold. We first compute the backtracking probability as a function of the shape of the inter-event time distribution. Second, we estimate the impact of the resulting bias to backtrack on the mixing rate of the process. Taken together, these results allow to quantify a mechanism that may either slow down or accelerate diffusion, by changing the number of steps leading to mixing. Such mechanism is inherently different from well-known mechanisms such as the bus paradox [Lambiotte et al. 2013] or other temporal mechanisms [Delvenne et al. 2015], only affecting the time to relaxation. Our observations also allow to gain insight on unexpected properties of a null model for temporal network analysis.

2.2 Passive edge-centric Random Walk on temporal network

2.2.1 Bias on the probability of backtracking

Let us first recall the passive edge-centric Random Walk as defined in section 1.5.3. Unless stated otherwise, we will simply call the passive edge-centric random as random walk (RW) through this chapter for the sake of notation. We consider a network where edges are activated for an infinitesimal duration. The time between two consecutive activations of an edge is governed by a renewal process, with i.i.d. inter-activation times distributed according to a probability density function. The activation processes on different edges are independent. For the sake of notation, we consider the same inter-activation distributions $f(t)$ for every edge; nonetheless the forthcom-

ing analytical derivations hold for distinct edge activities as well. The random walker waiting on a node jumps through the first edge incident to the node that is activated. The waiting-time distribution $g(t)$ on an edge ij, that is the time the walker arriving on the node i from an edge ki has to wait for the next activation of ij, is given by the bus paradox derived in equation (1.5.21):

$$g(t) = \frac{1}{\langle\tau\rangle}\int_t^{+\infty} f(\tau)\,\mathrm{d}\tau, \tag{2.2.1}$$

where $\langle\tau\rangle \equiv \int_0^{+\infty} \tau f(\tau)\,\mathrm{d}\tau$ is the mean inter-activation time.

It is important to note that this independence assumption is in general not respected if the walker passes several times through the same edge, as information about the previous passage time may help predicting the next activation time. This effect is most apparent for undirected networks, which we consider from now on. We will focus on such cycle effect on a directed network in Chapter 3, especially for cycles of length two. Consider a walker taking an edge from i to j. The waiting-time distribution on an edge jk is given by the bus paradox when k differs from i. However, this is not the case for the edge going back from j to i, i.e. for the backtracking transition, as illustrated in Figure 2.1. The waiting-time distribution for j to i is simply the inter-activation distribution $f(t)$, while we assume it is reasonably approximated by $g(t)$ for other edges. We neglect in particular memory effects due to the random walker exploring other longer cycles such as triangles, leading to a similar if attenuated effect.

The resulting statistical difference between backtracking and non-backtracking transitions may thus lead to biases in the dynamics of the walk, as the backtracking edge will be either favoured or penalized compared to the other edges depending on the underlying dynamics.

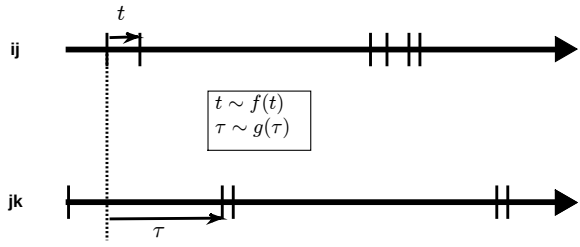

Figure 2.1 – Illustration of the backtracking bias on an edge ij. When the walker arrives on j via ij, the next activation time t of the edge ij is given by the distribution $f(t)$ associated to the renewal process on the edge ij, whereas the next activation time τ of another edge jk is given by the distribution $g(\tau)$ which is associated to the renewal process on the edge jk because of the paradox.

Let us now determine the probability $\mathscr{P}_d$ that the walker performs a backtracking jump as a function of the degree d of a node. Denoting $X_1,\ldots,X_{d-1} \sim g(t)$ the independent identically distributed waiting-times for the activation of the $d-1$ other competing edges, we have

$$\mathscr{P}_d = \int_0^{+\infty} P(r \leq \min_{k=1,\ldots,d-1} X_k) f(r)\,\mathrm{d}r\,. \tag{2.2.2}$$

Exploiting the independence of the edges and the equation (2.2.1) leads to

$$\begin{aligned}
P(r \leq \min_{k=1,\ldots,d-1} X_k) &= \prod_{k=1}^{d-1} \int_r^{+\infty} g_k(t)\,\mathrm{d}t && (2.2.3)\\
&= \left[\int_r^{+\infty} g(t)\,\mathrm{d}t\right]^{d-1} && (2.2.4)\\
&= \left[\int_r^{+\infty} \frac{1}{\langle\tau\rangle} \int_t^{+\infty} f(\tau)\,\mathrm{d}\tau\,\mathrm{d}t\right]^{d-1} && (2.2.5)\\
&= \left[\int_r^{+\infty} \frac{1}{\langle\tau\rangle} f(\tau) \int_r^{\tau} \mathrm{d}t\,\mathrm{d}\tau\right]^{d-1} && (2.2.6)\\
&= \left[\int_r^{+\infty} \frac{1}{\langle\tau\rangle} (\tau - r) f(\tau)\,\mathrm{d}\tau\right]^{d-1}, && (2.2.7)
\end{aligned}$$

where we have assumed that each edge has the same inter-activation distribution $g(t)$ for the sake of simplicity. When this assumption is not verified, the forthcoming development holds using equation (2.2.3) instead of (2.2.7).

Therefore, the expression of $\mathscr{P}_d$ is obtained as

$$\begin{aligned}
\mathscr{P}_d &= \int_0^{+\infty} \left[\int_r^{+\infty} \frac{1}{\langle\tau\rangle} (\tau - r) f(\tau)\,\mathrm{d}\tau\right]^{d-1} f(r)\,\mathrm{d}r && (2.2.8)\\
&= \int_0^{+\infty} \left[1 - \frac{\mathscr{F}(r)}{\langle\tau\rangle} + \frac{r}{\langle\tau\rangle}\,(F(r)-1)\right]^{d-1} f(r)\,\mathrm{d}r, && (2.2.9)
\end{aligned}$$

where $F(r) = \displaystyle\int_0^r f(\tau)\,\mathrm{d}\tau$ is the cumulative density function of $f(t)$ and appears explicitly , and $\mathscr{F}(r) = \displaystyle\int_0^r \tau f(\tau)\,\mathrm{d}\tau$.

The probability $\mathscr{P}_d$ depends on the number of competing edges $d-1$ but also on the shape of the distribution of the inter-event times. In particular, the presence of powers of r in the integral indicates that the shape of the inter-event time distribution impacts the backtracking probability $\mathscr{P}_d$ at least through its n first moments and thus through its variance. In the Poisson case, where $f(t)$ is an exponential distribution $\lambda e^{-\lambda t}$, the backtracking probability simplifies into the memoryless case $\mathscr{P}_d = \dfrac{1}{d}$ as expected.

Another interesting case is the power-law

$$f(\alpha,t_0) = \begin{cases} \dfrac{\alpha-1}{t_0}\left(\dfrac{t}{t_0}\right)^{-\alpha} & \text{if } t \geq t_0 \\ 0 & \text{otherwise,} \end{cases}$$

with $\alpha > 2$ (since the expression of $g(t)$ assumes finite mean), where

$$\mathscr{P}_d = \frac{(\alpha-1)^{2-d}}{(\alpha-2)d+1}. \tag{2.2.10}$$

Numerical simulations illustrate these results in Figure 2.2 where $f(t)$ follows various distributions including the exponential, gamma and power-law distributions. Note that the numerical convergence of the simulation is not guaranteed when the variance of the waiting-time distribution becomes infinite, which happens, for instance, for power-law distributions of exponent $\alpha < 3$. For each of these families of distributions, the higher the variance, the higher the probability of backtracking. However, as mentioned before, the backtracking probability depends, in general, on the full shape of the distribution.

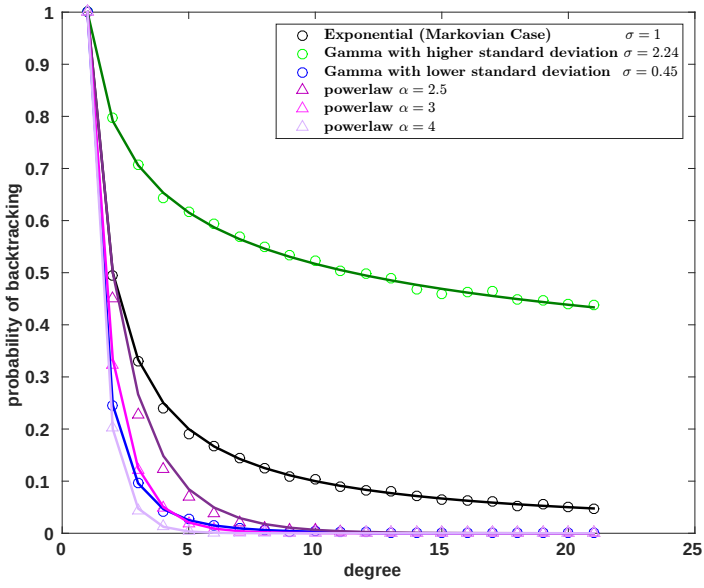

Figure 2.2 – Probability of backtracking from a node with respect to its out-degree for various distributions. Monte-Carlo simulations (circle) and theoretical curves obtained with equation (2.2.8) (solid line) coincide. For a given family, the higher the variance σ^2, the higher the probability of backtracking. For power-law distributions with small exponent, the backtracking probability remains large and decreases slowly as the degree increases.

As a next step, we test the importance of this effect in real-world systems by considering four datasets of face-to-face contacts described in [Génois et al. 2015; Gemmetto et al. 2014; Stehlé et al. 2011; Mastrandrea et al. 2015; Vanhems et al. 2013]. From the recorded contacts, we extract the largest connected component whose typical size is a few hundred nodes. We extract the inter-activation times between each pair of individuals, and aggregate them in the empirical inter-activation distribution $f(t)$. We simulate an RW on the corresponding homogeneous network where every edge activity is a renewal process governed by $f(t)$. The probability of backtracking as a function of the nodes degree, computed up to the largest node degree of the network , is displayed in Figure 2.3. Note that, similarly to the results of Figure 2.2, other structural properties of the network than the out-degree of the nodes have no impact on the probability of backtracking.

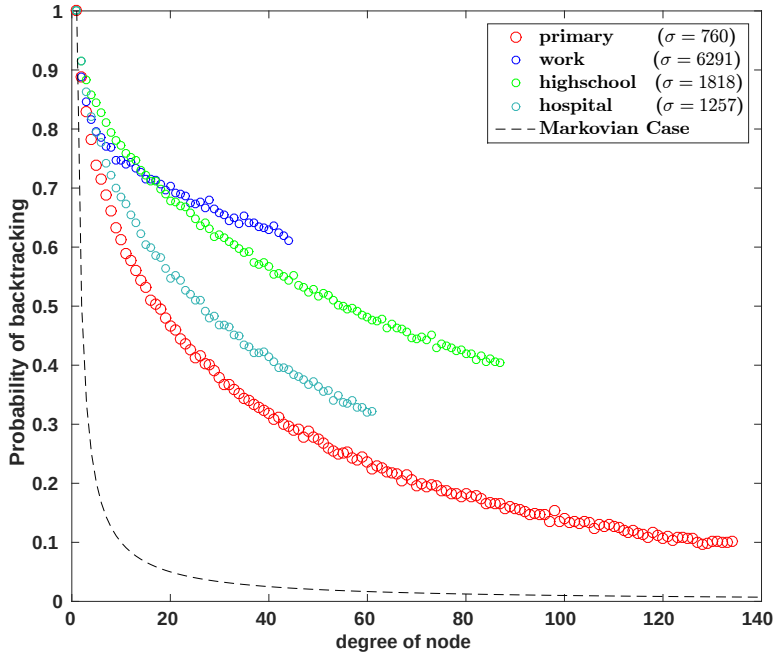

Figure 2.3 – Probability of backtracking from a node with respect to its out-degree for real-data. Each edge is governed by an i.i.d. renewal processes. All inter-event times on all edges have been aggregated to a unique global distribution $f(t)$, with standard deviation σ. The probability of backtracking is much larger than $\frac{1}{d}$ (corresponding to the Markovian case), even under destruction of the correlations between edges activities. For each dataset, the probabilities have been computed up to the largest nodes degree in the corresponding network.

We observe in the real-world data that backtracking is, overall, higher than in the memoryless case and increases with respect to the variance of the empirical inter-activation distribution. This is due to the fact the bus paradox has a larger impact when the variance increases. Intuitively, this implies that the consecutive face-to-face interactions between two persons tend be grouped or correlated (that is, A has several consecutive interactions with B followed by several interactions with C, rather than a mix of interactions with B and C at the same time).

This result shows that the backtracking bias is inherent to random walk processes, even when edge activities are uncorrelated. The paradox lies in the fact that the random walker has a tendency to take one particular edge over the others, even if each edge is statistically equivalent.

2.2.2 Impact of backtracking on the mixing rate of the Random Walk

In the previous section, we have shown that the shape of the inter-activation distribution may induce a backtracking bias for random walkers on a temporal network. We now estimate how this bias impacts the speed of diffusion, by estimating the mixing rate of the process.

From now on, we take a discrete-time perspective and no longer consider the timings at which events take place. The mixing time is measured by the number of steps k performed by the walker, thus focusing on the question: on average, how many steps are required for the process to reach equilibrium. This approach is in contrast with previous works focusing on the impact of the inter-activation distribution on the mixing time [Delvenne et al. 2015] and neglecting backtracking biases.

We first consider a standard memoryless RW process where no backtracking bias is present. In that case, the mixing rate is obtained from the second dominant eigenvalue of the transition matrix of the process, equivalent to the spectral gap of the corresponding normalized Laplacian of the graph. As we show in the appendix of this chapter, the spectral properties of the transition matrix are equivalent to those of a transition matrix defined on the so-called line graph, where edges of the original graph define nodes in the line graph. This equivalence is relatively intuitive, as both matrices essentially model the same process (only their representation changes), but it is crucial as a line graph formulation is natural to represent second-order Markov processes. In the following, we thus consider the spectral gap of the transition matrix of the line graph, defined as $1 - |\lambda_2|$, where λ_2 is the eigenvalue with the second largest module. The corresponding eigenmode describes the asymptotic dynamics of the process and is associated to the presence of bottlenecks/modules in the network [Delvenne et al. 2010]. The spectral gap provides information on the speed of convergence to stationarity since the distance between the transient state of the initial condition and the stationary state decays to 0 as $|\lambda_2|^k$ for large k. Therefore, the characteristic number of jumps for relaxation to stationarity is of the order $-\log|\lambda_2|$.

Table 2.1 compares the spectral gaps of the Markovian walker and the backtracking walker (with a backtracking probability computed using equation (2.2.8)) for four datasets, showing a significant slowdown of mixing due to backtracking alone. As a next step, we quantify this intuitive effect by performing a first-order approximation around the Markovian case, and focusing on regular networks for the sake of simplicity. Take an undirected network of ν nodes and μ edges, with an $\nu \times \nu$ adjacency matrix A and an incidence matrix K. From A, we get the stochastic transition matrices T of the network, and G_s of its line graph G associated to the standard memoryless RW. Importantly, it can be shown that both transition matrices T and G_s share the same non-zero eigenvalues (see appendix at the end of this chapter), which again may be intuitively understood as they correspond to the same linear dynamics described from the point of view of nodes and edges respectively.

	Spectral Gap of Markovian RW	**Spectral Gap of Backtracking RW**
Primary	0.4151	0.2738
Work	0.3057	0.0569
Highschool	0.1349	0.0396
Hospital	0.5695	0.2105

Table 2.1 – Shift of the spectral gap of the transition matrix due to the backtracking bias induced by the network temporality. The spectral gap is largely reduced, showing the strong impact of the inter-activity distribution on the number of steps required to explore the network.

We now consider a system where the trajectories of the walker are non-Markovian, such that the transition matrix M_s on the line graph differs from the transition matrix of the line graph associated to the Markovian case G_s. We consider a small deviation due to the probability of backtracking, by adding a perturbation matrix P :

$$M_s \quad = \quad G_s + P. \tag{2.2.11}$$

Each row of P captures the bias ε_{ji} of backtracking from the edge ij to the edge ji compared to a Markovian RW on T. The line of the matrix P corresponding to the jump transition from an edge ij is made of the entry ε_{ji} on the column corresponding to the edge ji, and $\dfrac{-\varepsilon_{ji}}{d(j)-1}$ for the $d(j)-1$ other edges leaving the node j, where $d(j)$ is the degree of node j. For the sake of simplicity, we calculate the impact of ε_{ji} on the spectrum of M_s for regular networks, where the degree is constant and the backtracking bias is thus $\varepsilon_{ji} = \varepsilon$ for every edge ji.
In this case, each eigenvalue λ of M_s is associated to an eigenvalue λ_0 of G_s through the equality $\lambda = \lambda_0 + \varepsilon\lambda^*$, where the perturbation λ^* is to be determined.

In the following, we perform standard derivations to obtain the first order approximation λ^* of the shift of the spectral gap.

The right eigenvector x_R of M_s associated to λ may also be expressed as $x_R = w_R + \varepsilon y_R$, where w_R is a right eigenvector of G_s associated to λ. We have

$$M_s x_R = (G_s + \varepsilon P)(w_R + \varepsilon y_R) \quad (2.2.12)$$
$$\lambda x_R = (\lambda_0 + \varepsilon\lambda^*)(w_R + \varepsilon y_R). \quad (2.2.13)$$

By definition of an eigenvector, equation (2.2.12) and (2.2.13) are equal.
Keeping the terms in ε and multiplying by the left eigenvector w_L^T of G_s associated to λ_0 yields to

$$w_L^T P w_R + w_L^T G_s y_R = \lambda^* w_L^T w_R + \lambda_0\, w_L^T y_R. \quad (2.2.14)$$

Isolating λ^* in Equation (2.2.14) leads to

$$\lambda^* \approx \frac{w_L^T P w_R}{w_L^T w_R} \quad (2.2.15)$$
$$= \frac{v_R^T K_{out} P K_{in}^T v_R}{v_R^T K_{out} K_{in}^T v_R} \quad (2.2.16)$$
$$= \frac{v_R^T K_{out} P K_{in}^T v_R}{\lambda_0 v_R^T D_A v_R}, \quad (2.2.17)$$

where K_{in} and K_{out} are two binary matrices such that the incidence matrix $K = K_{\text{in}} - K_{\text{out}}$.

The sign of the corresponding shift can be determined as follows. On the one hand, D_A is positive definite, on the other hand, $K_{out} P K_{in}^T$ is symmetric diagonally dominant with real non-negative diagonal entries by construction, hence positive semidefinite. Therefore, the sign of λ^* takes the sign of λ_0, and consequently the spectral gap of M_s decreases when the dynamics favours backtracking, and increases otherwise.

Finally, we validate our linear approximation of the real shift of the spectral gap $(1 - |\lambda_2|)$ with respect to the backtracking perturbation ε by computing the relative error on the spectral shift as

$$E_\varepsilon = \left| \frac{|\lambda_0 + \varepsilon\lambda^*| - |\lambda_2|}{1 - |\lambda_2|} \right| \quad (2.2.18)$$

on several regular graphs. As an illustration, Figure 2.4 shows the results for a regular network made of two communities of 50 nodes each, which is in the typical range of size of the four studied real-world networks (from 75 to 327 nodes). The approximation provides small relative error on the estimation of the spectral gap. Moreover, numerical simulations show that the linear approximation gives a lower bound to the true value of the spectral gap, and confirm the trend of slowing-down of the process under positive backtracking bias.

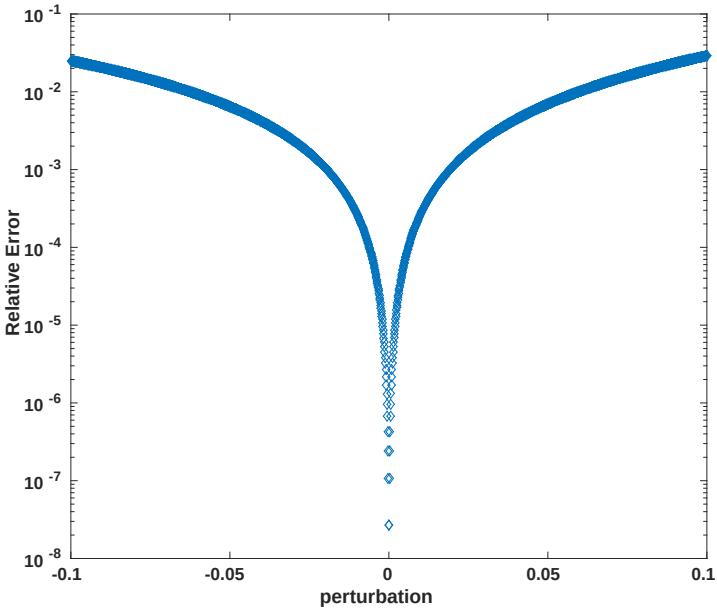

Figure 2.4 – Relative error E_ε of the linear approximation (Eq.2.2.18) in a network composed of 2 communities (2 cliques of 50 nodes of degree 50).

2.3 Discussion

The main purpose of this chapter was to highlight the existence of a neglected, yet important, correlation taking place in a null model actually designed to destroy temporal correlations in temporal networks. In models where edges are undirected and their activations are independent stochastic processes, dependencies between successive jumps of a passive random walker emerge, making the RW non-Markovian. Although we focused on backtracking in this chapter, it is clear that further memory is created in the RW by triangles, or short cycles in general. When the network is undirected, the backtracking bias is the dominating effect, as the exploration of a triangle by RW requires that each of the three consecutive steps is performed on a new edge of the triangle rather than backwards, which is unlikely when the backtracking bias is strong. On the contrary, backtracking is absent in directed networks (as a return edge does not necessarily exist or is ruled by an independent activation process) but the effects from short cycles remain. We will study in Chapter 4 the impact of the cycles of length two on an RW when the network is directed, in a larger context where edges remain active for non-infinitesimal durations.

Our findings question the relevance of standard models of diffusion on temporal networks: in the presence of bursty activation patterns, one cannot avoid correlations between events, either in the activation process or in the jumping process, making it a

non-trivial task to characterize the 'simplest' diffusion process with a given degree of burstiness. While we do know that real-world diffusion in social or mobility network exhibits non-Markovian patterns [Scholtes et al. 2014; Rosvall et al. 2014], those patterns sometimes favour and sometimes reject backtracking regardless of the degree of burstiness of the process, making it clear that they cannot be entirely accounted for by the effect at play in this paper. Whether the burstiness-induced memory is an undesirable artefact of the model or a useful and economical way to generate non-Markovian walks remains to be seen. On the one hand, it prevents from generating non-correlated RW. On the other hand, it allows to directly exhibit the impacts of the distributions involved in the process on the RW and leaves a signature that may allow one to differentiate between passive and active RW as well as between directed and undirected network, since the backtracking bias only exists for a passive RW on an undirected network. Moreover, the trajectory of an RW may be exploited to infer the degree of the nodes of a network based on the ratio of backtracking jumps at each node.

We have computed the effect of backtracking on the mixing rate of the diffusion process, due to its modification of the trajectories of the RW. This effect, however attenuated, holds for triangles or longer cycles, likely leading to a further asymptotic slowdown of the diffusion. This is a new mechanism for the impact of network temporality on diffusive processes which intrinsically differs from mechanisms such as the bus paradox and the consequential fact that the mixing time of bursty walker may be much larger than the naive estimate given by the number of jumps required to explore the network multiplied by the average waiting time of the walker at each step [Delvenne et al. 2015]. This is a tribute to the extraordinary richness of phenomena brought by the sole departure from a Poisson or discrete-time diffusion process.

Appendix of Chapter 2: Shared eigenvalues of the transition matrix and its associated transition line graph

We provide here a proof that the transition matrix T of a network and the one associated to its line graph G_s share the same non-zero eigenvalues.

We consider an undirected network of ν nodes and μ edges, with an $\nu \times \nu$ adjacency matrix A and its associated incidence $\nu \times 2\mu$ matrix K, listing each edge of the network in two consecutive columns of K with a $(+1,-1)$ entry and a $(-1,+1)$ entry for the two extremities of the edge (which extremity receives a $(+1,-1)$ or a $(-1,+1)$ being arbitrary). We decompose the incidence matrix into the difference of two binary matrices $K_{\text{in}} - K_{\text{out}}$.

First, it is direct that the adjacency matrix A and its associated line graph G share the same non-zero eigenvalues, since they are the commutated product of the same two rectangular matrices K_{in} and K_{out} :

$$\begin{aligned} A &= K_{\text{out}} K_{\text{in}}^T \\ G &= K_{\text{in}}^T K_{\text{out}} \end{aligned}$$

As a side note, G has at least $2\mu - \nu$ zeros eigenvalues, where ν and μ are respectively the number of nodes and edges of A. The transition matrices are obtained by normalizing the adjacency matrices by the degree of the nodes:

$$\begin{aligned} T &= D_A^{-1} K_{\text{out}} K_{\text{in}}^T \\ G_s &= {D_G}^{-1} K_{\text{in}}^T K_{\text{out}}, \end{aligned}$$

where D_A and D_G are the diagonal matrices of degrees of A and G respectively, and verify

$$D_G K_{\text{in}}^T = K_{\text{in}}^T D_A.$$

Let λ be a non-zero eigenvalue of T, and v an associated eigenvector of λ.
Then $w_R = K_{\text{in}}^T v$ is a right eigenvector of G_S associated to λ. Indeed:

$$\begin{aligned} & T v = \lambda v \\ \Leftrightarrow \quad & {D_A}^{-1} A v = \lambda v \\ \Rightarrow \quad & K_{\text{in}}^T K_{\text{out}} K_{\text{in}}^T v = \lambda K_{\text{in}}^T D_A v \\ \Leftrightarrow \quad & G K_{\text{in}}^T v = \lambda D_G K_{\text{in}}^T v \\ \Leftrightarrow \quad & G_s w_R = \lambda w_R. \end{aligned}$$

Similarly, the left eigenvector of G_s associated to λ is given by $w_L = K_{\text{out}}^T v$.

Chapter 3

Rock-Paper-Scissors Dynamics from Random Walks on Temporal Multiplex Networks

This chapter presents the results of [Gueuning et al.].

Abstract

We study diffusion on a multiplex network where the contact dynamics between the nodes is governed by a random process and where the inter-event time distribution differs for edges from different layers. We study the impact on an active edge-centric random walk of the competition that naturally emerges between the edges of the different layers. In opposition to previous studies, which have imposed a priori inter-layer competition, the competition is here induced by the heterogeneity of the activity on the different layers. We first study the precedence relation between different edges and by extension between different layers, and show that it determines biased paths for the walker. We also discuss the emergence of cyclic, rock-paper-scissors effects on random walks, when the precedence between layers is non-transitive. Finally, we numerically show the slowing-down effect due to the competition on a multiplex network with heterogeneous layers activity as the walker is likely to be trapped for a longer time either on a single layer, or on an oriented cycle.

3.1 Introduction

Random Walks have been notably studied in the context of multilayer temporal networks to uncover how the presence of multiple layers affects diffusion [Domenico et al. 2014] or to define generalized versions of Pagerank [De Domenico et al. 2015]. When studying edge-centric random walks, one typically assumes that the dynamics associated to the edges are either different for every edge (all-heterogeneous) [Masuda and Rocha 2018] or on the contrary are the same for every edge (all-homogeneous) [Masuda et al. 2017]. Here, we consider an intermediate situation where the edge activity depends solely on its layer: the edges in the same layer of the multiplex network have the same inter-event time distribution but these distributions may differ for edges in different layers. The system thus exhibits intra-layer homogeneity and inter-layer heterogeneity. This situation may be seen as an extension of the two typical frameworks as considering the existence of one single layer corresponds to the all-homogeneous case whereas considering one layer per edge corresponds to the all-heterogeneous case. We consider in particular noninterconnected (or edge-colored) multiplex networks. In such networks, the edges are associated to different layers and the nodes belong to each layer. As a consequence, there is no concept of switching between layers at a given node nor of inter-layer edges. Such model is appropriate to situations where a duplicate of each node is associated to each layer and where the time required to switch between layers at a node can be neglected.

In this chapter, we explore phenomena emerging from diffusion on temporal and multiplex networks. We study the properties of the resulting stochastic process and show that the presence of temporal heterogeneities across layers results in a competition between them and implies biases for the active edge-centric random walk trajectory, as defined in section 1.5.2. Here competition means that edges in one layer may have a higher probability to be selected by a walker than edges in another layer, due to the statistical properties of their temporal ordering. In other words, competition between layers emerges due to the temporality of the graph, and not as a model parameter as in previous works [Ding and Li 2017; Gómez-Gardeñes et al. 2015; Kleineberg and Boguñá 2016; Jang et al. 2015]. In addition, we show and explain some apparently counter-intuitive situations, such as the emergence of a cyclic, rock-paper-scissors precedence between the layers. Note that the notion of non-transitivity is well-known in statistics and that it has mostly focused on systems having a finite set of possible states, such as in non-transitive dice [Gardner 1970]. Our work can be seen as an extension to continuous variables and as a study of its impact on diffusion over multilayer networks. As a second step, we numerically explore the impact of the above mechanism on the dynamics of a walker and, specifically, we study the coverage of a walker on multiplex temporal network.

3.2 Active edge-centric Random Walk on multiplex temporal networks

3.2.1 Emergence of biased paths

We consider an active edge-centric RW, which means that when the walker arrives on a node, an inter-event time is associated to each edge leaving the node. Here, we do not consider the reduction of the model to an active node-centric RW because this reduction does not hold when the renewal processes associated to the edges are distinct. Indeed, the node-centric perspective assumes that all the out-going edges are statistically equivalent, whereas we will show later that this is not the case in the studied context. In particular, we will focus on the edges taken by the walker.

Let us consider the trajectory of a random walker. Arriving on node u, the walker the walker will leave through the first edge that reaches activation among the k edges connected to u. Because only the first edge to reach activation is taken by the walker, there is an underlying competition between the different edges.

Denote $T_1, \ldots, T_k$ the random variables associated to the inter-event times associated to each edge with distributions $f_1(t), \ldots, f_k(t)$. The transition time T of the walker, defined as the time before its next jump is given by:

$$T = \min(T_1, T_2, ..., T_k). \tag{3.2.1}$$

In our case, each edge belongs to a layer and it is thus natural to determine which layer is more likely to be selected by the walker. In particular, we will be interested in the notion of precedence between the random variables associated to the layers, which directly extends to the notion of precedence between layers. Formally, the nonnegative random variable A is said to precede the nonnegative random variable B, written $A \prec B$, if $P(A < B) > 0.5$, that is if for two edges starting from the nodes with inter-event times A and B respectively, the edge associated with A is more likely to reach activation before the edge associated with B than the opposite.

The existence of precedence relations translates into biased paths for the random walker jumping on the network since at each step of the walk some edges may be statistically more likely to be selected by the walker. Therefore, understanding the precedence relation between the random variables associated to the dynamics of the network is of paramount interest in order to understand the resulting diffusion.

In the following, we will exemplify the somewhat counter-intuitive properties of precedence, then we will numerically investigate its impact on the coverage of a random walker on a multiplex network with inter-layer heterogeneity in terms of activity, that is when the inter-event time distributions associated to the edges are different for edges of distinct layers.

3.2.2 Basic properties of precedence

3.2.2.1 Rock-Paper-Scissors

One may see precedence as a relation of dominance between edges in competition to attract the random walker. As such one may expect transitivity, that is if $A \prec B$ and $B \prec C$, one may expect $A \prec C$. However it is not the case, and we encounter circular, rock-paper-scissors situations as follows. We focus on the triangle network of Figure 3.2 and the three following distributions with expectation equal to 1 illustrated in Figure 3.1:

$$\begin{aligned} X &\sim \begin{cases} U[0.65, 0.75] & \text{with probability } 1 - \frac{1}{\varphi} \\ U[0.3\varphi + 0.65, 0.3\varphi + 0.75] & \text{with probability } \frac{1}{\varphi}; \end{cases} \\ Y &\sim U[0.9, 1.1]; \\ Z &\sim \begin{cases} U[0.75, 0.85] & \text{with probability } \frac{1}{\varphi} \\ U[0.2\varphi + 0.95, 0.2\varphi + 1.05] & \text{with probability } 1 - \frac{1}{\varphi}, \end{cases} \end{aligned} \tag{3.2.2}$$

where $U[a,b]$ stands for the uniform distribution on the interval $[a,b]$, and $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618$.

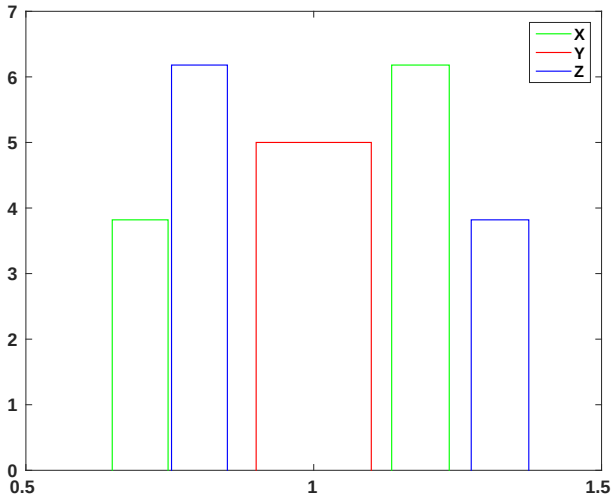

Figure 3.1 – The distributions associated to the random variables X, Y and Z of Equations 3.2.2.

It is straightforward to show that

$$P(Y < X) = P(Z < Y) = P(X < Z) = \frac{1}{\varphi}(\approx 0.61), \tag{3.2.3}$$

which consequently means that $Y \prec X$, $X \prec Z$ and $Z \prec Y$. As a consequence, a walker jumping on the network will have a tendency to jump clockwise on the triangle, as illustrated numerically in Figure 3.2. The emergence of a circular flow reflects correlations between the successive edges on the random walk trajectory, even though edges are chosen independently at random at each step.

There is a parallel that may be made with the phenomenon of Brownian or Feynman-Smoluchowski ratchet theory where fluctuations or noise may induce work [Feynman et al. 1965; Astumian 1997]. It considers a microscopic motor corresponding to a wheel with a ratchet that rotates freely clockwise but is prevented to rotate anti-clockwise. The impulse to the wheel is provided by the Brownian motion of a particle and the ratchet's motion may provide work to an external system. Here the clockwise trajectory of the walker corresponds to the rotation of the wheel and originates from the particular distributions associated to the different edges.

Competition becomes more complex when more than two edges interact together. For instance, in the above example, even if the pairwise precedence is uniform, one finds as $P(X < \min(Y,Z)) = 1 - \frac{1}{\varphi} > \frac{1}{3}$ when considering the competition between 3 different edges, and thus X will tend to be favoured by the walker against the other two in the presence of the three types of edges.

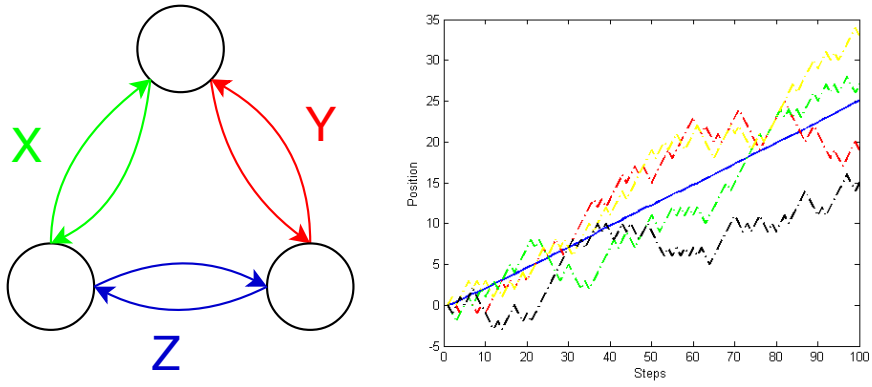

Figure 3.2 – Numerical simulations of an RW on a multiplex triangle [left]. Each (colored) pairwise relation belongs to a different layer and distributions activity of the layers are given by the distribution of (3.2.2). The walker position is incremented $+1$ at each clockwise jump, and -1 at each anti-clockwise jump. On the right hand side, the four dash-dot lines represent four independent RW starting at 0 while the blue line represents the average over 1000 independent RW. The precedence relation is not transitive as an RW jumping on the triangle will have a tendency to jump clockwise.

3.2.2.2 There is no most preceding edge

It is clear that comparing either means or variances of two random variables is not sufficient to determine which one precedes the other. Even more, it is impossible to find a random variable that precedes any random variable with equal mean. Without loss of generality, we prove this statement for unit-mean random variables.

First, we observe that for any random variable X following distribution $f(t)$, it is possible to find a random variable Y_n of distribution $g_n(t)$ such that Y_n precedes X, by setting

$$g_n(t) = \begin{cases} \frac{n}{2}(n-1) & \text{if } t \in \left[\frac{1}{n} - \frac{1}{n^2}\ ,\ \frac{1}{n} + \frac{1}{n^2}\right] \\ \frac{1}{2\sqrt{n}} & \text{if } t \in \left[n-1+\frac{1}{n} - \frac{1}{\sqrt{n}}\ ,\ n-1+\frac{1}{n} + \frac{1}{\sqrt{n}}\right], \\ 0 & \text{otherwise} \end{cases} \tag{3.2.4}$$

where $n \in \mathbb{N}_0$ is chosen large enough. One possible choice of n is such $n \geq 3$ and such that $\frac{1}{n} + \frac{1}{n^2} < \beta$, where β is the 10% quantile of f. Indeed, in this case one finds

$$\begin{aligned} P(Y_n < X) &> P(Y_n < \beta \cap \beta < X) && (3.2.5)\\ &= P(Y_n < \beta) \times P(\beta < X) && (3.2.6)\\ &> P\left(Y_n < \frac{1}{n} + \frac{1}{n^2}\right) \times 0.9 && (3.2.7)\\ &= \frac{n}{2}(n-1)\left(\frac{1}{n} + \frac{1}{n^2} - (\frac{1}{n} - \frac{1}{n^2})\right) \times 0.9 && (3.2.8)\\ &= \left(1 - \frac{1}{n}\right) \times 0.9 && (3.2.9)\\ &> 0.5, && (3.2.10) \end{aligned}$$

which implies that Y_n precedes X.

From the sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ with distribution respectively given by $(g_n)_{n\in\mathbb{N}}$, it is possible to extract a subsequence of random variables $(Z_n)_{n\in\mathbb{N}}$ whose respective distributions have non-overlapping supports, for instance by defining $Z_n = Y_{10n}\ \forall n \in \mathbb{N}$. Such a sequence is increasingly precedent, that is, for any $n \in \mathbb{N}$ we have $Z_{n+1} \prec Z_n$.

As the series $(g_n)_{n\in\mathbb{N}}$ does not converge to a probability distribution, there exists no most preceding random variable because any arbitrary random variable is preceded by a random variable from $(Z_n)_{n\in\mathbb{N}}$. In terms of diffusion on multiplex networks, this result implies that there is no optimal a priori distribution ensuring that a given layer always captures a majority of the RW flow, independently of the distributions in the other layers. However, once the inter-event time distributions are associated to the other layers, a layer may always find a distribution that will allow it to be the most precedent.

3.2.2.3 Layer precedence at a node and node out-degree

The notion of precedence between random variables naturally extends to precedence between layers when inter-event time distributions inside a layer are homogeneous but vary across layers. In this case, the layer L_1 precedes the layer L_2 at node u if the walker sitting on u is more likely to perform the next jump through an edge of L_1, that is if L_1 is more likely to capture the flow passing through u. Moreover, in the presence of several layers, L_1 precedes L_2 and L_3 jointly at node u if L_1 precedes the artificial layer L_2L_3 at node u, which is obtained by the union of the layers L_2 and L_3 .

Precedence of layers at a node depends on its out-degree k_i on each layer L_i, as the layer of the edge selected by a walker is determined by comparing the smallest time m_i to reach activation on each layer L_i, where m_i is the minimum of k_i random variables with identical distributions associated to the layer L_i. In the simplest case of two duplicate layers L_1 and L_2 with inter-event times on edges of the type X and Y respectively, that is L_1 and L_2 share the same nodes and have the same edges, the out-going edges consist in k edges of each type, and L_1 precedes L_2 at node u if

$$\min_{j=1,\dots,k} X_j \prec \min_{j=1\dots,k} Y_j, \tag{3.2.11}$$

where X_j and Y_j are duplicates of random variables following the same distribution than X and Y respectively.

The layer precedence at a node may therefore vary between nodes depending on their out-degree, and in particular between high and low out-degree nodes. However, there always exists a threshold out-degree ν^* above which one layer will always precede the other one. Indeed, let the distribution f be said to have a larger minimal weight than the distribution g if there exists $\varepsilon > 0$ such that $P(X < \varepsilon) > P(Y < \varepsilon)$ and $\forall\ 0 < \sigma < \varepsilon$, $P(X < \sigma) \geq P(Y < \sigma)$ where X and Y are two random variables following the distribution f and g respectively. Then, there always exists a threshold out-degree ν^* above which the minimum of at least ν^* realizations of the distribution with the larger minimal weight will precede the other one.

In order to investigate this effect in a real-world setting, we construct a multiplex network as follows. We use a dataset of private messages sent on an online social network at the University of California Irvine [Pietro et al. 2009]. This typical social network is then duplicated to create an hypothetical two-layer social network, where each layer can be thought of as corresponding to a different medium of communication. Each layer is thus identical and composed of 1899 nodes and 20296 edges. The difference between the layers is induced by the choice of two different distributions of X and Y defined in Eq (3.2.2), where $Y \prec X$. As X has a larger minimal weight than Y, there exists a switch in the precedence at a node between the variables of type X and Y depending on the out-degree of the node. In this case, it is straightforward to show that this switch occurs when at least two edges of each type are competing, that is

$$\min_{i=1,\dots,\nu} X_i \prec \min_{i=1\dots,\nu} Y_i \ \ \forall \nu \geq 2. \tag{3.2.12}$$

Figure 3.3 shows the probability of taking an edge of type X instead of Y with respect to the out-degree of a node in the static aggregated network (by construction, twice the out-degree in each layer), where each edge has two activation times associated to the random variables X and Y respectively. The numerical results confirm a switch of the precedence when the out-degree of the nodes increases.

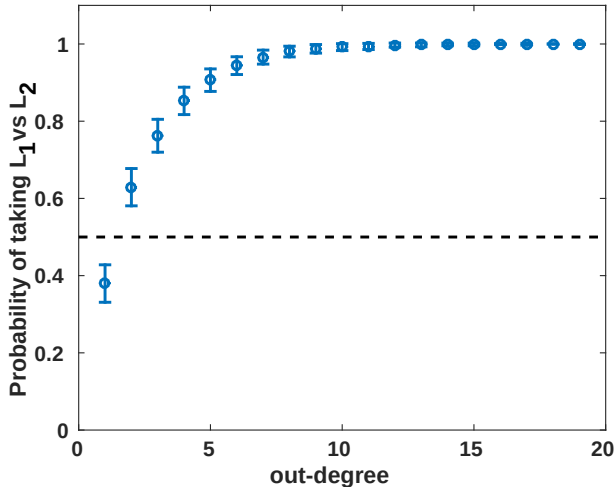

Figure 3.3 – Average probability ($\pm$ the standard deviation) of taking an edge of L_1 versus one of L_2, with respect to the out-degree of each node on the replicate layers, over 10000 simulations. A switch in the layers precedence occurs between the nodes of out-degree equal to one and the nodes of out-degree ≥ 2. The distribution of the inter-event times of the edges of L_1 and L_2 are the distributions of the random variables X and Y respectively, defined in Eq (3.2.2). Dashed line corresponds to a probability equal to 0.5.

3.3 Impact of inter-layer activity heterogeneity on the coverage of the Walker

Finally, we investigate numerically the impact of competition in the case of diffusion in a multiplex social network with more than two layers. To do so, we use publicly available data introduced in [Magnani et al. 2013] where the layers consist in five kinds of social relationships between 61 employees. The corresponding network has in total 620 undirected edges. In the numerical simulations, we consider the three types of random variables defined in Equation (3.2.2) and focus on the layers associated to Facebook, Leisure and Co-authorship relations which have 193, 124 and 88 edges respectively. Unlike the previous simulations, the existence of an edge on one

layer does not imply that such an edge exists on every layer, the layer associated to Facebook being for instance denser than the other ones. The structure of the graph thus corresponds to real-world interactions, while the inter-events time are chosen for the sake of illustration. Indeed, the presence of three layers and the choice of the specific inter-event time distributions allow the emergence of properties of precedence we have previously shown, such as a rock-paper-scissors situation, and thus enables the investigation of its impact on the random walk. We tested all the possible associations of a specific distribution from equation (3.2.2) to a given layer and the results remained consistent and similar to the ones presented in Figure 3.4.

We numerically compute the coverage $\rho(k)$ of a walker, as defined in section 1.5.5 as the percentage of distinct nodes visited by the walker after k steps, and use Equation (1.5.38) as theoretical prediction of the coverage on a undirected interconnected multiplex network for the homogeneous case:

$$\rho(k) \quad = \quad 1 - \frac{1}{N^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \exp(-\mathbf{F_j} \mathbb{T}_\nu \mathbf{E_i}^T). \tag{3.3.1}$$

Here, we always assume homogeneous inter-event time distribution inside a given layer, and consider homogeneous as well as heterogeneous inter-layer inter-event time distributions. Again, we assume that the edges belonging to the same layer are governed by the same inter-event time distribution. Inter-layer homogeneity corresponds to situations where the distributions associated to the different layers are identical, and inter-layer heterogeneity to the situations where these distributions are distinct.

A typical simulation is provided in Figure 3.4 (left), where we observe that the coverage of a random walk tends to grow faster under inter-layer homogeneity compared to inter-layer heterogeneity, irrespective of the choice of the unique inter-event time distribution. The slow-down induced by the inter-layer heterogeneity is mainly due to the fact that the flow is captured inside a single layer. Since the random walker has a tendency to stay in this layer, the walk will mainly take place on this layer, that is less connected than the aggregated network. However, when the graph density of the layers is weak, a rock-paper-scissors situation may arise, as the switch in precedence between layers might not occur for lower out-degree nodes. This emergence promotes the switch of the walker between different layers through lower out-degree nodes, resulting in a more efficient exploratory walk across the network. We illustrate this effect in Figure 3.4 (right) through an artificial multiplex network, where each node has one outgoing edge in each layer, pointing to a randomly chosen node at each layer, ensuring low density and homogeneous out-degree distribution. Rock-paper-scissors situations occur at each step of the walk because there is a competition between one edge of each type at each node. Such situations impact the trajectory of the walker as it prevents the flow to be captured into a single layer. Thus, the trajectory of the walker will be similar to the one under inter-layer homogeneity. Therefore, the coverage of the walk is in this case similar to the coverage of an homogeneous inter-layer network (or monolayer network).

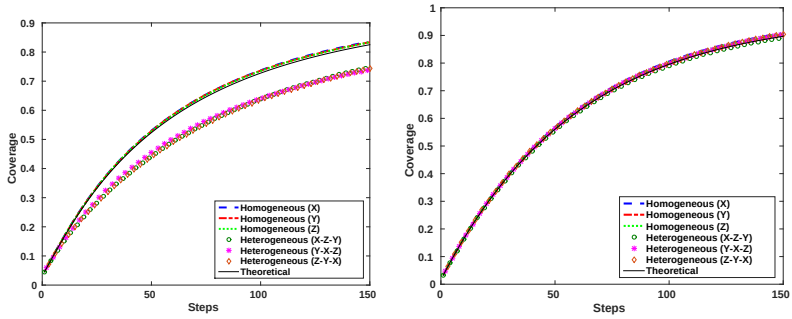

Figure 3.4 – Coverage over 100 paths starting at each node of the multiplex network. Left: real-world network with layers corresponding to Facebook, Leisure and Co-authorship relations respectively; Right: artificial network where the out-degree of each node on each layer is set to one. The distributions X, Y and Z come from Eq.(3.2.2), and are assigned to the different layers. Different combinations are possible and we distinguish inter-layer homogeneous situations, e.g. Homogeneous (X), from heterogeneous situations, where each layer has its own inter-event time distribution, e.g. Heterogeneous (X,Y,Z). The coverage is larger under homogeneous inter-layer activity (dotted lines) than under heterogeneous inter-layer activity (markers) [left] because the layer preceding the others two tends to capture the flow, except when out-degrees are low [right]. Numerical simulations agree with standard theoretical predictions (solid line) in the homogeneous case.

3.4 Discussion

The main purpose of this work was to investigate the competition between different layers of a multiplex network in situations where the network is temporal. In our framework, edges activations are modeled as independent renewal processes, each layer being characterized by a different inter-event time distribution, and we highlight the implications of the concept of precedence on diffusion. In particular, we have shown that precedence may lead to biases between the different layers of the network. Despite the simplicity of the process, it may lead to counter-intuitive properties, such as non-transitivity, out-degree dependence, or rock-paper-scissors situation. Our numerical results show that precedence may have important quantitative effects on the speed of diffusion on a multiplex network, as the precedence of one layer over others may hinder the number of edges available to the walker, and hence slow down its coverage of the graph. We also show that high out-degree nodes are more prone to favor one single layer, while low out-degree nodes exhibit a different effect and may lead to a cyclic exploration between the layers. We have studied the impact of the precedence on an active edge-centric RW, however it is worth noticing that its impact on the passive RW may lead to opposite bias towards layers due to an additional competition induced by the short cycles as we will study in chapter 4. For non-exponential distri-

butions, this effect results in a backtracking bias towards or against the last traveled edges as we have shown in chapter 2, typically leading to the emergence of short cycle patterns in human-related network [Saramäki and Holme 2015].

These results remain mostly mathematical as we used toy-model distributions instead of ones modelled on real-life data. An important next step would be to test the resulting ideas on empirical data of multiplex networks where different layers are associated to different time scales, for instance between physical, mobile phone and social media interactions [Sekara et al. 2016].

Chapter 4

Emergence of Memory in Random walk on temporal networks with lasting edges

This chapter presents the main results of [Petit et al. 2018].

Abstract

We consider a random walk that is both node-centric and edge-centric on a directed temporal network where the edges remain activated for non-infinitesimal duration. The dynamics is governed by three types of independent stochastic processes related to the walker's waiting time, and to the up-times and down-times of the edges. We first study the trajectory of a random walker on a directed acyclic graph, then extend the study to graph with cycles. We study how the walker's trajectory is affected by the emerging memory induced by the short cycles, regardless to the Markovian nature of the underlying stochastic processes in play. In particular, we characterize the impact of cycles of length two on the trajectory of the walker, while the method naturally extends to longer cycles.

4.1 Introduction

The majority of the literature of diffusion on temporal network assumes that edges are activated for an infinitesimal duration, as the time scale for an event duration is assumed to be much smaller than the one of the diffusing entity. However this assumption does not always hold, as it has been observed in real-life datasets [Gauvin et al. 2013; Zhao et al. 2011; Scherrer et al. 2008].

Taking into account these finite durations has been shown to lead to practical implications, for instance in community detection [Sekara et al. 2016]. Our objective is to study how taking into account non-vanishing edges affects a random walker trajectory. We focus on directed network, hence the random walker's trajectory is not affected by a backtracking bias, which we studied in chapter 2. Here, the dynamics is governed by three types of stochastic processes associated to: i) the walker's waiting-time on a node, ii) the duration of the activation of an edge and iii) the time between two consecutive activations of an edge.

In this chapter, we first derive a master equation for the walker on a Directed Acyclic Graph (DAG) and show how some limiting cases correspond to results already known in the literature. Indeed, the problem is quite rich as three different timescales are involved, and we show that our framework includes the standard active node-centric and passive edge-centric random walks introduced in sections 1.5.1 and 1.5.3 respectively. Then, we consider networks with cycles and provide a correction to the obtained equations in order to take into account the impact of cycles. In particular, we focus on the correction due to cycles of length two, as their impact is the strongest, however the method holds for longer cycles as well.

4.2 Active node-centric and passive edge-centric Random Walk on temporal network with lasting edges

4.2.1 Model description

We consider a random walk on a directed temporal network where the edges remain activated for non-infinitesimal durations. The duration of the activation of an edge ij is governed by a renewal process, with i.i.d times distributed according to an Up-time probability density function $U_{ij}(t)$. Similarly, the time between two consecutive activations of an edge ij is governed by a renewal process, with i.i.d. times distributed according to a down-time probability density function D_{ij}, which indicates how long the edge ij remains inactive. Thus, the activity of an edge ij is described by the alternation of up, or activated, states of duration drawn from U_{ij} followed by down, or deactivated, states of duration drawn from D_{ij}, as illustrated in Figure 4.1. The renewal processes associated to the edges are independent.

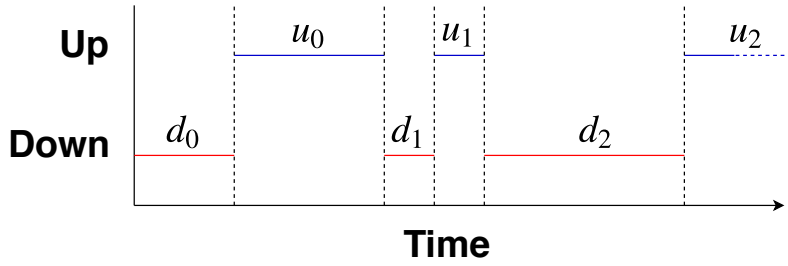

Figure 4.1 – Illustration of the activity of an edge. An edge ij alternates between active/up phases (in blue) of random duration $u_k \sim U_{ij}(t)$ and inactive/down phases (in red) of random duration $d_k \sim D_{ij}(t)$.

The random walk is defined as follows and illustrated in Figure 4.2. The walker is jumping on a directed network that evolves over time. The underlying network is associated to the adjacency matrix A which encodes all the possible connections between nodes in the network. When arriving on a node i, the walker is first assigned a waiting time t_w according to a waiting-time distribution $\psi_i(t_w)$. This time t_w corresponds to the minimal time the walker will spend on a node before performing a new jump. One may picture t_w as the time required for the walker to visit the node or to regain enough energy to be able to perform a new jump. Once this minimal waiting-time t_w is elapsed, the walker is ready to jump through one outgoing edge leaving from i that is activated. If several edges are available, the walker selects one of them with uniform probability. If on the contrary no edge is available, the walker is trapped on the node i and stays there until one out-going edge from i is finally activated, allowing the walker to perform a jump. As we consider continuous-time distributions, a trapped walker will not have to make a choice between two distinct edges, as the probability that two edges are activated simultaneously is almost surely zero.

It is worth mentioning that [Figueiredo et al. 2012] studied a similar walk with the distinction that a trapped random walker is assigned a new waiting-time from the same distribution $\psi_i(t_w)$, and that the authors only focused on the asymptotic state of the process.

4.3 Discussion on the different timescales

The dynamics of the random walk depends on the three timescales associated to the speed of the walker, the up-times and the down-times of the edges. When the timescale of the walker significantly differs from the one associated to the edges, the context of the walk may be simplified, as depicted in Figure 4.3.

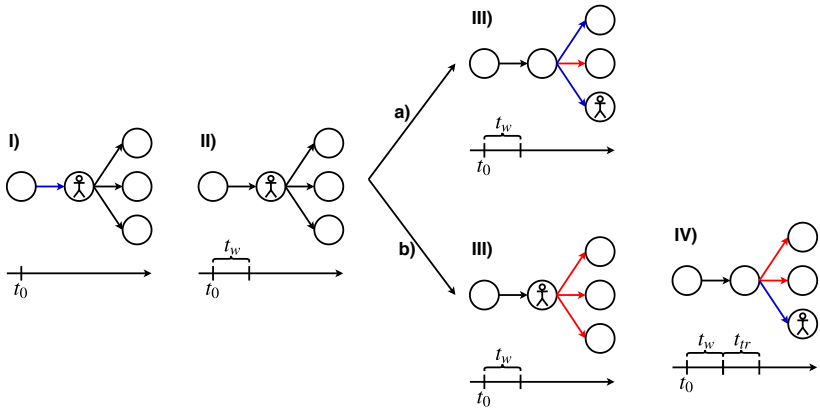

Figure 4.2 – Illustration of the Random Walk. Edges are colored in blue when they are up, in red when they are down, and in black when their current state does not impact the walk. Arriving on a node i (step **I)**) at time t_0, the walker first waits for a minimal waiting-time t_w (step **II)**). Once the duration t_w elapses, the walker is ready to jump. Two situations may happen: a) either at least one out-going edge from the node i is activated, and the walker performs a jump after randomly selecting one of them with equal probabilities (step **III)**); b) or no out-going edge from i is activated (step **III)**), and the walker is trapped on the node i until one out-going edge ij is finally activated, through which the jumper performs a jump (step **IV)**).

Taking the timescale λ of the walker as reference, four extreme situations may arise:

A: **Low down-times and high up-times.** The global underlying network is almost always available to the walker, as edges are sporadically down for small durations. Thus, the walker will never be trapped. Therefore, only the dynamics of the walker impacts the process and the walk corresponds to an active node-centric random walk on the underlying static network.

B: **Low down-times and low up-times.** The network is rapidly "blinking" as edges switch quickly between up and down states. Thus, the walker will never be trapped for a non-negligible duration. The situation is therefore comparable to the previous case, and the walk corresponds to an active node-centric random walk on the underlying static network. When the up-times increase, the network blinks more slowly and the active node-centric model holds.

C: **High down-times and high up-Times.** The network smoothly changes as from time to time one edge is (de)activated. However, the walker may perform a significant number of jumps between any two consecutive changes in the topology. Thus, the walk corresponds to consecutive active node-centric random walks on

successive subgraphs of the global underlying networks. Therefore, the modelling through a node-centric random walk on the global underlying network fails to model the dynamics, as it over-estimates the density of the network on which the walker is jumping. When the down-times decrease, the consecutive realizations of the network tend to be more similar to the underlying network, and the active node-centric model becomes a better proxy of the process.

D: **High down-times and low up-times.** The network appears really sparse to the network, as some edges are sporadically activated for small duration. Thus, the network corresponds to a temporal network with infinitesimal durations in the walker's perspective. Therefore, the walker is always trapped and his initial waiting-time arriving on a node is negligible. Therefore the model corresponds to a passive edge-centric random walk. When the up-times increase or the down-times decrease (or both simultaneously), the interplay between the different dynamics becomes complex, and there is no standard model for these intermediate states.

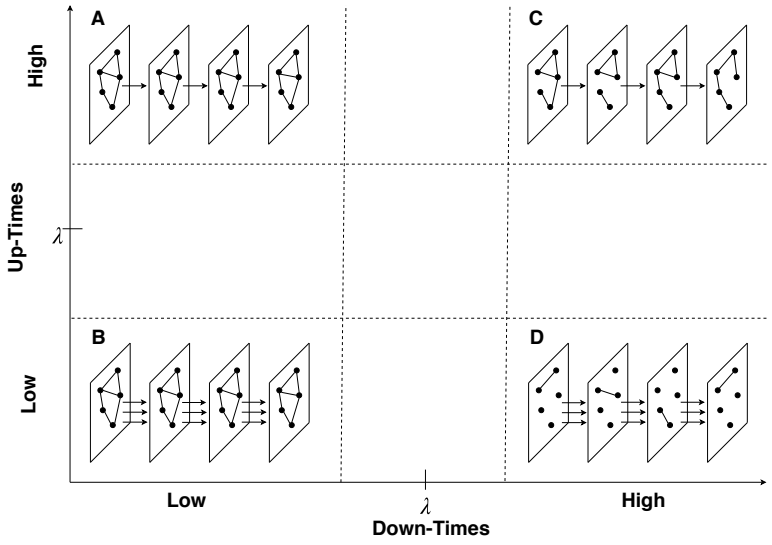

Figure 4.3 – Variations of the model with respect to the relative characteristic timescales of the up-times and down-times of the edges compared to the timescale λ of the walker . The walker sees the network as consecutive subnetworks of the underlying adjacency matrix that switch with different speed according to the timescales. Single arrow (resp. triple arrows) between layers indicate slow (resp. fast) switches between layers, in the walker's perspective. When the timescales are well separated, standard active node-centric (from **A** to **B**) or passive edge-centric (**D**) models capture the dynamics. Our model fills the gap when the standard models do not hold (between $\{$**A,B**$\}$ and **D**).

As we mentioned, when the three timescales are not significantly distinct (central region of the Figure 4.3), or in some cases when two of the timescales are of the same order, both the dynamics of the active node-centric random walk and of the passive edge-centric random walk fail to describe the true dynamics. This situation is the focus of the following work. As a motivating illustration, we consider the random walk on a toy example network depicted in Figure 4.4 and vary the rates of the up-times and down-times that both follow exponential distributions.

The evolution of the density $n(t)$ of the random walk is obtained through numerical simulations. The density of the corresponding active node-centric n_{active} and the passive edge-centric n_{passive} random walks are obtained by solving numerically equations (1.5.1) and (1.5.16) respectively. We compute the norm of the error for a simulation duration T between the numerical simulation and the two model as

$$E_{\text{model}}(T) \quad = \quad \int_0^T \left\| \mathbf{n}_{\text{model}}(s) - \mathbf{n}(s) \right\|_2 \mathrm{d}s, \tag{4.3.1}$$

where "model" stands for "active" or "passive".

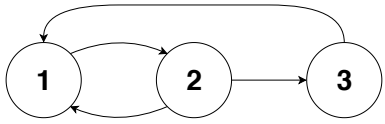

Figure 4.4 – Toy-Model network used for simulations of Figure 4.5

Consistently with the schematic representation of Figure 4.3, the errors E_{active} and $E_{passive}$ shown in Figure 4.5 indicate that the standard models hold when timescales are significantly distinct but fail to cover the full domain of timescales. Therefore, even the all-exponential distributions situation is not yet covered by the classical models.

One of the reasons of this failure is that the underlying network does not exactly correspond to the one that is available to the walker, as the connection listed in the adjacency matrix of the underlying network are not always available to the walker. A more fundamental reason lies in the fact that the standard models assume independence between the successive activations of the edges. However, the history of the walker trajectory may influence its future when the walker visits the same node in a short time, and similarly to the backtracking bias we explored in chapter 2, memory may naturally emerge when cycles exist, even though the dynamics associated to the walker and to the edges are Markovian.

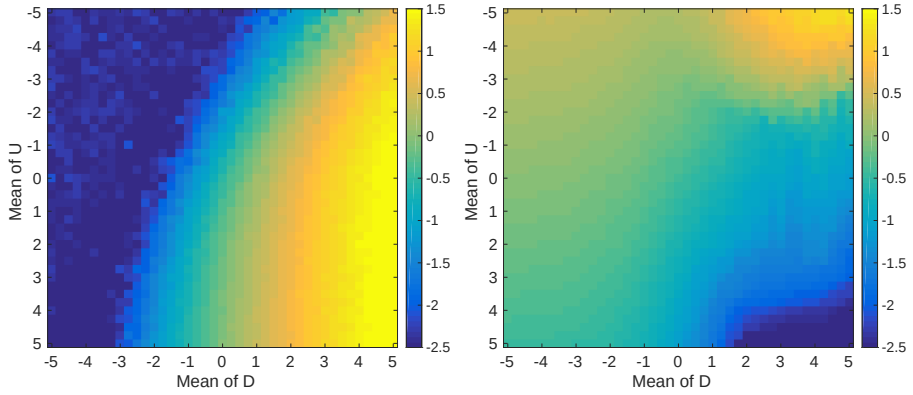

Figure 4.5 – Logarithmic error of the active node-centric (left) and the passive edge-centric (right) computed from Equation 4.3.1. The models fail to capture the dynamics as schematically illustrated in Figure 4.3.

In the remaining part of this chapter, we present a general framework that holds for not well-separated timescales on a directed acyclic graph, and provide a correction that allows to take into account the impact of the cycles of length two on the trajectory.

4.4 Dynamics on a Directed Acyclic Graph

As a first step, let us consider the trajectory of the walker on a Directed Acyclic Graph (DAG). As a DAG is acyclic, emergence of memory due to cycles is prevented. Hence the approximation on a DAG for a more general network will allow to characterize the impact of memory in a second step. It is worth mentioning that DAGs have many applications [Melnik et al. 2011] and allow coarse-grain model of any directed graph, through the mapping of each strongly connected component of the initial directed graph to a single node of the DAG.

In the case of a DAG, the random walk may be considered as a passive node-centric Random Walk as discussed at section 1.5.4. However we will present another approach that allows a natural extension to cyclic graphs.

4.4.1 Master equation on a DAG

Similarly to Chapter 1, let $n_i(t)$ denote the density of the walker at node i. Denoting $q_i(t)$ the probability density function of the arrival time on node i and $\Phi_i(t,\tau)$ the probability to stay on node i on the interval $[\tau,t]$, with τ the arrival time on node i, the

walker's density is obtained through

$$n_i(t) = \int_0^t q_i(\tau)\Phi_i(t,\tau)\mathrm{d}\tau. \tag{4.4.1}$$

The probability $\Phi_i(t,\tau)$ depends on the probability density distributions $T_{ji}(t,\tau)$ of the transition time from i to j, where the walker arrives on node i at time τ:

$$\Phi_i(t,\tau) \quad = 1 - \int_\tau^t \sum_{j\in V_i} T_{ji}(\nu,\tau)\mathrm{d}\nu. \tag{4.4.2}$$

Applying Leibniz's rules to deal with the integral when differentiating, one obtains the rate of evolution of the walker on the node i as

$$\dot{n}_i(t) \quad = q_i(t) - \int_0^t q_i(\tau) \sum_{j\in V_i} T_{ji}(t,\tau)\mathrm{d}\tau. \tag{4.4.3}$$

Defining the diagonal integral operator by its i^{th} component as

$$(\mathscr{D}\mathbf{q}(t))_i = \int_0^t q_i(\tau) \sum_{j\in V_i} T_{ji}(\nu,\tau)\mathrm{d}\tau, \tag{4.4.4}$$

equation (4.4.3) can be written in vectorial form as

$$\dot{\mathbf{n}}(t) \quad = (I - \mathscr{D})\mathbf{q}(t). \tag{4.4.5}$$

Now, we want to find an expression of $\mathbf{q}(t)$ that only depends on the transition densities T_{ji} and on the initial condition $\mathbf{n}(0)$.

Let $q_i^{(k)}(t)$ be the probability to arrive on node i at time t in exactly k jumps. The initial condition directly provides $q_i^{(0)}(t) = n_i(0)\delta(t)$, where $\delta(t)$ stands for the dirac function. For $k > 0$ we have:

$$\begin{aligned}
q_i(t) &= \sum_{k=0}^{\infty} q_i^{(k)}(t) && (4.4.6)\\
&= \sum_{k=0}^{\infty} q_i^{(k+1)}(t) + q_i^{(0)}(t) && (4.4.7)\\
&= \sum_{k=0}^{\infty}\sum_j \int_0^t q_j^{(k)}(\nu)T_{ij}(t,\nu)\mathrm{d}\nu + \mathrm{q_i^{(0)}(t)} && (4.4.8)\\
&= \sum_j \int_0^t \sum_{k=0}^{\infty} q_j^{(k)}(\nu)T_{ij}(t,\nu)\mathrm{d}\nu + \mathrm{q_i^{(0)}(t)} && (4.4.9)\\
&= \sum_j \int_0^t q_j(\nu)T_{ij}(t,\nu)\mathrm{d}\nu + \mathrm{q_i^{(0)}(t)}, && (4.4.10)
\end{aligned}$$

where equation (4.4.6) allows the last simplification to equation (4.4.10).
Defining the linear integral operator $\mathscr{T}$ as

$$\mathscr{T}\mathbf{q}(t) = \int_0^t T(t,\nu)\mathbf{q}(\nu)\mathrm{d}\nu, \tag{4.4.11}$$

where $T(t,v)$ is a matrix function with component (i,j) given by $T_{ij}(t,v)$, allows to write equation (4.4.10) in a vectorial form:

$$\begin{aligned}
\mathbf{q}(t) &= \mathscr{T}\mathbf{q}(t)+\mathbf{q}^{(0)}(t) && (4.4.12)\\
&= (I-\mathscr{T})^{-1}\mathbf{q}^{(0)}(t) && (4.4.13)\\
&= \sum_{k=0}^{\infty}\mathscr{T}^k\mathbf{q}^{(0)}(t) && (4.4.14)\\
&= \sum_{k=0}^{\infty}\mathscr{T}^k\delta(t)\mathbf{n}(0), && (4.4.15)
\end{aligned}$$

where the expression of $(I-\mathscr{T})^{-1}$ follows from the Neumann's lemma, and is valid because $\mathscr{T}$ is linearly bounded since the equation 4.4.10 is a Volterra integral equation of the second kind.
Therefore, the master equation (4.4.5) can be re-written in terms of the transition densities and the initial condition:

$$\dot{\mathbf{n}}(t) = (I-\mathscr{D})\sum_{k=0}^{\infty}\mathscr{T}^k\mathbf{q}^{(0)}(t) \quad (4.4.16)$$

4.4.2 Transition density on a DAG

For the master equation (4.4.16) to be tractable, one needs to explicit $T_{ji}(t,\tau)$ in terms of the model parameters (up-time, down-time and waiting-time distributions along with the underlying structure). For the sake of simplicity, we will now consider that all nodes are associated with the same waiting-time $\psi(t)$, and that all edges have the same up-time distributions $U(t)$ and down-time distributions $D(t)$, that is $\psi_i(t)=\psi(t)$, $U_{ij}(t)=U(t)$ and $D_{ij}(t)=D(t)$ for all $i,j=1,\ldots,n$. Nonetheless the forthcoming analytical derivations hold for distinct edge activities as well.

Let $p(t)$ denote the probability that a given edge ij is up at a random time t. With no other information on the history of the activations of the edge, we have that $p(t)$ is constant, and given by

$$p(t) = \frac{\langle U\rangle}{\langle U\rangle+\langle D\rangle} := p, \quad (4.4.17)$$

where the symbol $\langle.\rangle$ again stands for the expectation of the distributions.

The transition density from i to j can be decomposed into a sum of two terms, $T_{ji}(t,\tau)$ **=(1) + (2)**, depending on whether the walker is trapped or may directly jump once he is ready.

The first term **(1)** corresponds to the situation where the walker is not trapped, which occurs when edge ij is up when the walker is ready. For the edge to be selected by the walker without being trapped, several independent conditions need to be respected simultaneously:

1. The waiting-time of the walker must be exactly $t-\tau$: $\psi(t-\tau)$;
2. The edge ij must be up: p ;
3. the edge ij must be selected among all the other edges thar are also up at time t: $\sum_{k=1}^{|V_i|} \frac{1}{k}\binom{|V_i|-1}{k-1} p^{k-1}(1-p)^{|V_i|-k}$, where $|V_i|$ is the out-degree ot node i.

Further computation of the product of the three terms provides

$$(\mathbf{1}) \quad = \quad \psi(t-\tau)\frac{1}{|V_i|}\left[1-(1-p)^{|V_i|}\right]. \tag{4.4.18}$$

The second term corresponds to the situation where the walker was trapped at a time before t, and edge ij is the first out-going edge to be activated, exactly at time t. The probability for an edge to be down when the walker is ready to jump is given by $1-p$. The situation where the walker has to wait for an edge ij to be activated corresponds to the situation leading to the bus paradox (explained in section 1.5.3). Therefore, the time the walker has to wait before the edge ij is activated follows the probability density function provided by

$$\mathscr{D}(t) = \frac{1}{\langle D\rangle}\int_t^\infty D(v)\mathrm{d}v. \tag{4.4.19}$$

To determine the term **(2)**, one needs to consider each waiting time $x-\tau$ of the walker which is smaller than $t-\tau$ and weight by:

1. The probability that all edges are down at time $t-x$: $(1-p)^{|V_i|}$;
2. The fact that exactly one edge will be up after an extra waiting of $t-x$: $\mathscr{D}(t-x)$:
3. The probability that all the other edges will remain down for a time longer than $t-x$: $\left[\int_{t-x}^\infty \mathscr{D}(s)\mathrm{d}s\right]^{|V_i|-1}$;

which results to

$$(\mathbf{2}) \quad = \quad (1-p)^{|V_i|}\int_\tau^t \psi_i(x-\tau)\mathscr{D}(t-x)\left[\int_{t-x}^\infty \mathscr{D}(s)\mathrm{d}s\right]^{|V_i|-1}\mathrm{d}x. \tag{4.4.20}$$

In short, we have shown that

$$T_{ji}(t,\tau) = c_1\psi_i(t-\tau) + c_2\int_\tau^t \psi_i(x-\tau)\left[\int_{t-x}^\infty \mathscr{D}(s)\mathrm{d}s\right]^{|V_i|-1}\mathscr{D}(t-x)\mathrm{d}x, \tag{4.4.21}$$

where the constant c_1 and c_2 only depend on $|V_i|, \langle U\rangle$ and $\langle D\rangle$.
Interestingly, the full shape of the down-time distribution D matters, whereas only the mean of the up-time distribution U matters. The reason is that when the edge is up, the jump occurs instantaneously whereas the total duration of the down-time period impacts the walker's trajectory.

4.4.3 Limiting cases with exponential distributions on a DAG

Before considering graphs with cycles, we consider some limiting cases on a DAG and show that we recover standard results.

4.4.3.1 Only down-times

Let us first consider the situation when edges are always down and the waiting-times of the walker are infinitesimal, that is $D(t)$ is arbitrary and $U(t) = \psi(t) = \delta(t)$. The minimum of $|V_i|$ independent random variables with density $\mathscr{D}$ follows

$$\mathscr{D}_{(1),i}(t) \quad = |V_i| \left(\int_t^\infty \mathscr{D}(s)\mathrm{d}s\right)^{|V_i|-1} \mathscr{D}(t), \tag{4.4.22}$$

Thus, equation (4.4.21) simplifies as

$$T_{ji}(t,\tau) \quad = \quad \frac{1}{|V_i|}\mathscr{D}_{(1),i}(t-\tau), \tag{4.4.23}$$

and we recover the dynamics of an active edge-centric random walk (see section 1.5.1).

4.4.3.2 Only down-times and waiting-times

Let us consider the previous situation and add non-negligible waiting-time distributions for the walker, that is $D(t)$ and $\psi(t)$ are arbitrary and $U(t) = \delta(t)$.
Then equation (4.4.21) simplifies as

$$\begin{aligned}
T_{ji}(t,\tau) &= \int_\tau^t \psi(x-\tau)\mathscr{D}(t-x)\left[\int_{t-x}^\infty \mathscr{D}(s)\mathrm{d}s\right]^{|V_i|-1}\mathrm{d}x && (4.4.24)\\
&= \frac{1}{|V_i|}\int_\tau^t \psi(x-\tau)\mathscr{D}_{(1),i}(t-x)\mathrm{d}x && (4.4.25)\\
&= \frac{1}{|V_i|}\int_0^{t-\tau} \psi(y)\mathscr{D}_{(1),i}(t-\tau-y)\mathrm{d}y && (4.4.26)\\
&= \frac{1}{|V_i|}\int_0^{+\infty} \psi(y)\mathscr{D}_{(1),i}(t-\tau-y)\mathrm{d}y && (4.4.27)\\
&= \frac{1}{|V_i|}\left(\psi * \mathscr{D}_{(1),i}\right)(t-\tau) && (4.4.28)
\end{aligned}$$

and we recover the dynamics of an active random walk which is both node-centric and edge-centric, which we have shown to be equivalent to a single active node-centric random walk (see section 1.5.4).

4.5 Dynamics on directed networks with cycles

The previous results rely on the assumption that the states (up or down) of the successive edges taken by the walker are independent. However, this assumption fails under the presence of cycles. Indeed, when the walker arrives on a node i already visited, the states of each outgoing edge from i at the time of the previous visit of the walker at the node i are known by the walker, and may therefore bias the walker's trajectory, in a similar fashion than the backtracking effect shown in chapter 2. The prediction of the acyclic model may still make good predictions when the walker's dynamics governs the process, or when the out-degree are large, as bias effects are diluted; on the other cases, significant deviations between the true dynamics and the DAG approximation of the model can be observed even when all dynamics are Markovian as we discussed in section 4.3.

When the underlying network contains cycles, the full trajectory needs to be considered in the state space for the model to be exact, which is not analytically tractable. We will therefore focus on a generalized method taking into account the cycles of length two, which have the stronger effect. The method can be extended to longer cycles.

4.5.1 Generalized master equation with correction for cycles of length two

We need to enlarge the state space of the system in order to allow a correction for cycles of length two. As a consequence, we will extend the memory to the last two steps: the walker sitting on node i has now the memory of having jumped from m' to m at time ν, and from m to i at time τ, as depicted in Figure 4.6.

With this memory, let us define:

- the arrival time density $q_{imm'}(\tau,\nu)$ for the couple of times (τ,ν);
- the conditional transition density $T_{j|imm'}(t|\tau,\nu)$ of jumping from i to j at time t;
- the probability $\Phi_{imm'}(t|\tau,\nu)$ to stay until time t on node i.

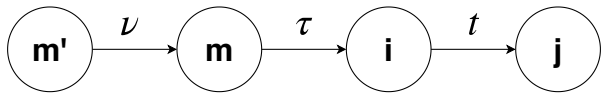

Figure 4.6 – Two-steps memory of the walker: jump from m' to m occurred at time ν and from m to i at time τ. The goal is to determine the transition probability of jumping to j at time t. The edges are labeled with the time of the jump.

We have

$$\Phi_{imm'}(t|\tau,\nu) = 1 - \sum_{j\in V_i} \int_\tau^t T_{j|imm'}(s|\tau,\nu)\,\mathrm{d}s, \tag{4.5.1}$$

with, because the walker will never stop jumping,

$$\sum_{j\in V_i} \int_\tau^\infty T_{j|imm'}(s|\tau,\nu)\,\mathrm{d}s = 1 \tag{4.5.2}$$

for all $0 \le \nu \le \tau$ and $1 \le i \le N$.

Following the same steps than for the DAG, we will determine the density of the walker in function of the initial condition $\mathbf{n}(0)$, taking into account the cycles of length two.
The first two steps of the walk are not impacted by the memory. Thus, the probability $n_i(t)$ that the walker is on node i at time t decomposes into

$$n_i(t) = n_i^{(0)}(t) + n_i^{(1)}(t) + n_i^{(k\ge 2)}(t), \tag{4.5.3}$$

where the superscripts refer to the number of jumps performed up to time t, and $n_i^{(k\ge2)}(t) = \sum_{k\ge2} n_i^{(k)}(t)$. The first two terms are not impacted by the memory effect and can be computed based on the transition densities established in the DAG case:

$$n_i^{(0)}(t) = n_i(0)\Phi_i(t,0), \tag{4.5.4}$$

and

$$n_i^{(1)}(t) = \int_0^t q_i^{(1)}(\tau)\Phi_i(t,\tau)\mathrm{d}\tau \tag{4.5.5}$$

$$= \sum_{m\in V_i'} n_m(0) \int_0^t T_{im}(\tau,0)\Phi_i(t,\tau)\mathrm{d}\tau. \tag{4.5.6}$$

For all $k \ge 2$, $n_i^{(k)}(t)$ is given by

$$n_i^{(k)}(t) = \sum_{m'\to m\to i} \sum \iint_{0\le\nu\le\tau} q_{imm'}^{(k,k-1)}(\tau,\nu) \tag{4.5.7}$$

$$\times \Phi_{imm'}(t|\tau,\nu)\,\mathrm{d}\nu\mathrm{d}\tau, \tag{4.5.8}$$

where again the superscript in $q_{imm'}^{(k,k-1)}$ gives the number of jumps that have been made before performing the jumps on the path imm'.
Therefore, to determine $n_i^{(k\ge2)}(t)$ we only need

$$q_{imm'}(\tau,\nu) = \sum_{k\ge2} q_{imm'}^{(k,k-1)}(\tau,\nu). \tag{4.5.9}$$

Let us determine the arrival times density in a given number of jumps, $q^{(k,k-1)}_{imm'}(\cdot,\cdot)$. Again, the first two steps are memoryless, thus $q_{imm'}(\tau,\nu)$ may be decomposed as

$$q_{imm'}(\tau,\nu) = q^{(2,1)}_{imm'}(\tau,\nu) + \sum_{k=2}^{\infty} q^{(k+1,k)}_{imm'}(\tau,\nu). \tag{4.5.10}$$

Due to the lack of memory for the first two steps, the initial condition of arrival times after the first two steps is determine using the transition density of the acyclic case $\widetilde{T}_{ji}(t) := T_{ji}(t,0)$. The walker must initially be at node m', jump to m at time ν and jump to i after sitting exactly $t-\tau$ at time m:

$$q^{(2,1)}_{imm'}(\tau,\nu) \quad = \quad \widetilde{T}_{im}(\tau-\nu)\widetilde{T}_{mm'}(\nu)n_{m'}(0) \tag{4.5.11}$$

When at least two steps have been performed, that is for all $k \geq 2$, we have

$$q^{(k+1,k)}_{imm'}(\tau,\nu) \quad = \quad \sum_{m''\in V'_{m'}} \int_0^{\nu} T_{i|mm'm''}(\tau|\nu,\nu') q^{(k,k-1)}_{mm'm''}(\nu,\nu')\,\mathrm{d}\nu' \tag{4.5.12}$$

Summing all the terms leads to

$$q_{imm'}(\tau,\nu) = \sum_{m''\in V'_{m'}} \int_0^{\nu} T_{i|mm'm''}(\tau|\nu,\nu') q_{mm'm''}(\nu,\nu')\,\mathrm{d}\nu' + q^{(2,1)}_{imm'}(\tau,\nu). \tag{4.5.13}$$

One may incorporate equation (4.5.13) into equation (4.5.7) and then into equation (4.5.3), and the density of the walker is obtained in function of the initial condition and the transition densities $T_{j|imm'}(t|\tau,\nu)$, which will be determined in the following.

4.5.2 Transition density with correction for cycles of length two

We want to compute $T_{j|imm'}(t|\tau,\nu)$. The trajectory before the jump at time ν is not taken into account and so only durations starting from time ν matter, which means that we can shift all the times by ν :

$$\begin{aligned} T_{j|imm'}(t|\tau,\nu) &= T_{j|imm'}(t-\nu|\tau-\nu,0) && (4.5.14)\\ &:= \widetilde{T}_{j|imm'}(t-\nu|\tau-\nu) && (4.5.15)\\ &= \widetilde{T}_{j|imm'}(x|y), && (4.5.16) \end{aligned}$$

with $x = t-\nu$ and $y = \tau-\nu$.
Given the previous jumps from m' to m then from m to i, three situations are distinguishable when considering a new jump from i to j, as illustrated in Figure 4.7 :

1. $m' \neq i$: The walker did not take a cycle of length two (Figure 4.7 I)), thus no memory effect is in play and the model for DAGs hold. Hence

$$\widetilde{T}_{j|imm'}(x|y) \quad = \quad \widetilde{T}_{ji}(x-y) \tag{4.5.17}$$

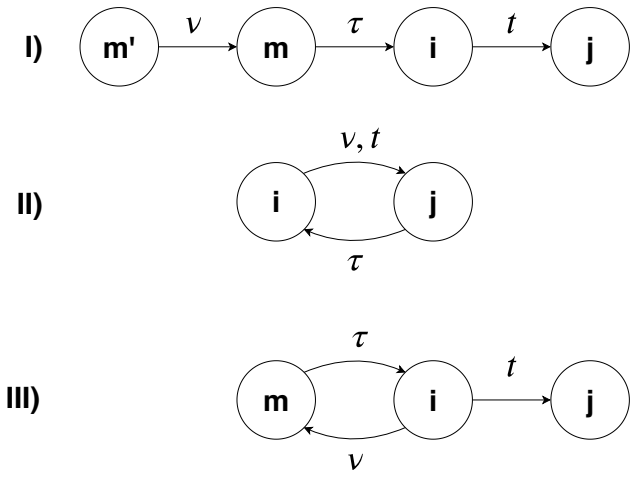

Figure 4.7 – The three possible situations for a jump with two-steps memory: I) no cycle; II) continuing on a cycle; III) leaving a cycle. The edges are labeled by the time of the jump through it.

2. $m' = i$ and $m = j$: The walker just jumped through a cycle of length two, and performs his next jump through the cycle (Figure 4.7 II)).

3. $m' = i$ and $m \neq j$: The walker just jumped through a cycle of length two, and performs his next jump through an edge that does not belong to the cycle (Figure 4.7 III)).

The second and third cases are impacted by the memory, but the corresponding transition density can still be decomposed into the sum of two terms, depending on whether the walker is trapped or may directly jump once he is ready:

$$\widetilde{T}_{j|imm'}(x|y) \quad = \quad (\mathbf{1}) + (\mathbf{2}). \tag{4.5.18}$$

Before computing the two terms of the sum, one needs to compute the probability for an edge to be up, given that the decision whether to jump across it has been taken at a previous time or not.

4.5.2.1 Corrections on the probability p for an edge to be up

Let us assume that the walker completed a cycle of length two.

When considering the random walk on a DAG, one assumed the knowledge of no other information on the history of the activations of the edges. Consequently, the probability p for an edge to be up was constant and equal to $\frac{\langle U \rangle}{\langle U \rangle + \langle D \rangle}$. This formula does not hold in the case of cycles of length two because a competition between all

out-going edges from i already happened at time τ. Thus, we know that at least the edge ij was up at time τ. Therefore the probability for ij to be up a time t is not $\frac{\langle U\rangle}{\langle U\rangle+\langle D\rangle}$ anymore. This probability also changes for all the other edges ij' that were not selected. The deviation from p is stronger as $t-\tau$ is smaller, and typically decreases over time to reach the value of $\frac{\langle U\rangle}{\langle U\rangle+\langle D\rangle}$. Therefore, one needs to discriminate between the probability associated to the edge ij, and the ones associated to the other edges, and we will denote

$$p_i^*(s,\nu) = \mathrm{P}\{ij \text{ is up at time } s \,|\, \text{jumped across it at time } \nu\}, \quad (4.5.19)$$

$$p_i^\dagger(s,\nu) = \mathrm{P}\{ij' \text{ is up at time } s \,|\, \text{jumped across } ij \text{ at } \nu\} \quad (4.5.20)$$

for some $s \geq \nu$ and $j' \neq j$.
We will present the derivations of $p_i^*(s,\nu)$, the derivations for $p_i^\dagger(s,\nu)$ are provided in the appendix of the reference paper [Petit et al. 2018].

Let us define $\tilde{q}_i$, the probability that the walker was trapped before performing its first jump through ij. As we discard memory beyond the last two jumps, the trapping occurs only if all the $|V_i|$ edges were down when the walker was ready to jump, which is simply given by

$$\tilde{q}_i = (1-p)^{|V_i|}. \quad (4.5.21)$$

Let us consider the density of the extra up-time the edge ij remained up after the first jump of the walker. If the walker was trapped, then the extra up-time follows the density $U(t)$ as the walker jumped at the beginning of an active phase of the edge ij; if the walker was not trapped, he arrived at a random time over an active phase of the edge ij, and the corresponding extra up-time follows the distribution $\mathscr{U}(x)$ which is given by the bus paradox equation (1.5.21) applied to $U(x)$. Weighting with $\tilde{q}_i$, the global density $\widetilde{U}_i(x)$ of the extra up-time is given by

$$\widetilde{U}_i(x) = \tilde{q}_i U(x) + (1-\tilde{q}_i)\mathscr{U}(x). \quad (4.5.22)$$

The edge ij is up at time s if it stayed always up from τ to s, or if it switched its state an even number of time on the interval (ν,s). The time for an edge to switch twice its stat is given by the convolution of the distributions U and D because it is just the sum of the random variables following these distributions. Thus, considering all the possible cases, one finds

$$\begin{aligned} p_i^*(s,\nu) &= \int_{s-\nu}^{\infty} \widetilde{U}_i(r)\mathrm{d}r \\ &\quad + \sum_{k=0}^{\infty} \int_0^{s-\nu} \left(\widetilde{U}_i * D^{*(k+1)} * U^{*k}\right)(r) \int_{(s-\nu)-r}^{\infty} U(t)\,\mathrm{d}t\,\mathrm{d}r, \end{aligned} \quad (4.5.23)$$

where the superscript $*k$ denotes k auto-convolution of the density function.

Now that p is corrected to take into account cycles of length two, we may develop the terms of the equation (4.5.18).

4.5.2.2 Second term of the transition density

Let us focus on the situation where the edge ij belongs to the previously visited cycle of length two.
The equations correspond to the equations (4.4.18) and (4.4.20) where p has been replaced accordingly by $p^*(t,v)$ and $p^\dagger(t,v)$, which provide:

$$\begin{aligned}
(\mathbf{1})_{(j|iji)} &= \psi_i(t-\tau)\sum_{k=1}^{|V_i|}\frac{1}{k}p_i^*(t,v)\binom{|V_i|-1}{k-1} \\
&\quad \times (p_i^\dagger(t,v))^{k-1}(1-p_i^\dagger(t,v))^{|V_i|-k} \qquad (4.5.24)\\
&= \frac{p^*(t,v)}{p_i^\dagger(t,v)}\psi_i(t-\tau)\left[\frac{1-(1-p_i^\dagger(t,v))^{|V_i|}}{|V_i|}\right]. \qquad (4.5.25)\\
(\mathbf{2})_{(j|iji)} &= \int_\tau^t \psi_i(s-\tau)(1-p_i^*(s,v))\mathscr{D}(t-s) \\
&\quad \times\left[(1-p_i^\dagger(s,v))\left[\int_{t-s}^\infty \mathscr{D}(r)\mathrm{d}r\right]\right]^{|V_i|-1}\mathrm{d}s. \qquad (4.5.26)
\end{aligned}$$

4.5.2.3 Third term of the transition density

Let us focus on the situation where the edge ij does not belong to the previously visited cycle of length two.
The second term is still given by equation (4.5.26), hence only the first term requires further development.

The transition density for ij to be chosen a time t is given by the product between $\psi_i(t-\tau)$ and the probability for the walker to choose the edge ij. Such probability can be developed by considering two mutually exclusive situations: whether the edge im belonging to the cycle is up or down.

If im is up, which happens with probability $p_i^*(t,v)$, then there is a competition between the edge ij, the edge im and k other edges, where k may vary between 0 and $|V_i|-2$:

$$\begin{aligned}
\mathrm{P}\{\text{choose } ij \mid im \text{ is up}\} &= p_i^\dagger(t,v)\sum_{k=0}^{|V_i|-2}\binom{|V_i|-2}{k} \\
&\quad \times\frac{1}{k+2}(p_i^\dagger(t,v))^k(1-p_i^\dagger(t,v))^{|V_i|-k-2} \\
&= \frac{|V_i|p_i^\dagger(t,v)+(1-p_i^\dagger(t,v))^{|V_i|-1}}{|V_i|(|V_i|-1)p_i^\dagger(t,v)} \qquad (4.5.27)
\end{aligned}$$

If im is down, which happens with probability $1-p_i^*(t,\nu)$, then there is a competition between the edge ij and k other edges, where k may vary between 0 and $|V_i|-2$:

$$\begin{aligned} \mathrm{P}\{\text{ choose } ij \mid im \text{ is down }\} &= p_i^\dagger(t,\nu)\sum_{k=0}^{|V_i|-2}\binom{|V_i|-2}{k}\frac{1}{k+1} \\ &\quad\times(p_i^\dagger(t,\nu))^k(1-p_i^\dagger(t,\nu))^{|V_i|-k-2} \quad (4.5.28) \\ &= \frac{1-(1-p_i^\dagger(t,\nu))^{|V_i|-1}}{|V_i|-1}, \quad (4.5.29) \end{aligned}$$

The final form of the first term is therefore obtained by

$$\begin{aligned} (\mathbf{1})_{(j|imi)} &= \psi_i(t-\tau)\times\Big[p_i^*(t,\nu)\mathrm{P}\{\text{choose } ij \mid im \text{ is up}\} \\ &\quad +(1-p_i^*(t,\nu))P\{\text{choose } ij \mid im \text{ is down}\}\Big], \quad (4.5.30) \end{aligned}$$

where the probabilities in curly braces are provided by equations (4.5.27) and (4.5.29).

We stick to this analytical approach and indicate to the interested readers that numerical simulations validating the correction associated to cycles of length two are provided in [Petit et al. 2018] on toy examples, when all the distributions associated to the process are exponentials and independent, as even in this situation the Markovian properties of the walk are lost and one needs to take into account the cycles' corrections.

4.6 Discussion

The main purpose of this chapter was to take into account the non-infinitesimal durations of the edges when considering diffusion on temporal networks and to estimate how this impacts the dynamics of a passive RW. We have highlighted the importance of three timescales to characterize the diffusion on temporal networks: one related to the diffusive entity *on* the network and two related to the dynamics *associated to* the network itself.

Our main contribution was to exhibit deviations in the trajectory of a passive RW when the short cycle-induced memory is neglected and to provide corrections in order to predict the trajectory of the RW. We focused on deriving the analytical density of the walker. We have shown that our expressions are exact for DAGs and developed corrections due to cycles of length two when the underlying network admits cycles.

The developed method is general and includes standard models when one or several timescales can be neglected. Moreover, the model is flexible and one may extend it to take into account corrections due to longer cycles, according to the particular structure of the network of interest and the associated occurrence of cycles of particular lengths.

The work is mainly theoretical but it has a lot of potential applications. For instance, many collected face-to-face interaction datasets contain contacts' duration, like the SocioPatterns data [Gemmetto et al. 2014; Génois et al. 2015] , or data collected via Bluetooth [Sekara et al. 2016]. Other practical applications concern networks of mobile sensors that can only share information for limited periods, typically depending on the proximity between the agents. As RW are used in many applications, our results may lead to generalization of several standard tools.

Part III

Spreading strategies on temporal networks

Chapter 5

Imperfect spreading on temporal networks

This chapter presents the results of [Gueuning et al. 2015].

Abstract

We study the impact of non-Markovian activity on spreading on networks. We consider a process of imperfect spreading where transmission is successful with a determined probability at each contact. We first derive an expression for the inter-success time distribution, determining the speed of the propagation, and then focus on a problem related to epidemic spreading by estimating the epidemic threshold in a system where nodes remain infectious during a finite, random period of time. Finally, we discuss the implications of our work on designing an efficient strategy to enhance spreading on temporal networks.

5.1 Introduction

We now investigate the impact of bursty behaviour on a spreading process taking place on a temporal network. As we mentioned in Section 1.6, a spreading process differs from a diffusion process in the sense that there is no conservation of the quantity of the moving entity. A spreading process based on a Random Walk's perspective consists in the jumper leaving a replica of itself at each jump. Therefore, there is no emergence of competition between the edges leaving from the same node, as the events on one edge do not affect events occurring on the other one. As a consequence, when considering a classical SI model, only the speed of the diffusion is affected by bursty/non-Markovian behaviour.

In this chapter, we investigate a model where spreading is not always successful on an edge when it appears, so that it takes place only after a random number of trials. This process allows us to introduce, and to tune, different timescales in the system, one associated to the network evolution through its inter-event times, and the other to the spreading process through the transmission probabilities. As we show analytically and numerically, imperfect spreading leads to a generalization of the bus paradox we encountered previously and has interesting implications on the properties of epidemic spreading on networks. In particular, we focus on the way to improve the speed of a spreading process when the spreader has to choose between tuning the transmission probabilities or the inter-event rates.

5.2 Imperfect spreading on temporal networks

5.2.1 Estimating the inter-success time

The spreading model is defined as follows, and illustrated at Figure 5.1. We consider a temporal network where the activation of edges is governed by a renewal process, with inter-event distribution $f(\tau)$. Again, it is assumed that edges have a vanishing duration, and that the random processes associated to the different edges are independent. Spreading takes place from node to node. The diffusing entity, called the virus from now on for the sake of simplicity, sits on a node until an edge appears. The virus then invades the neighbouring node with a probability p. As long as the neighbouring node is not infected, the virus will try to invade it at each activation of the edge with the same probability p. Following the taxonomy of Chapter 1, the process is therefore passive edge-centric. In order to describe the random process, it is crucial to estimate the waiting time before a certain edge is invaded, which we call the inter-success time. So far, this model is essentially equivalent to an SI model [Wearing et al. 2005], which we assume to take place on a tree in order to avoid non-linear effects.

In situations when $p = 1$, the relation between the inter-event distribution and the inter-success distribution $\varphi(t)$ is given by the bus paradox. Indeed, as the activations of different edges are independent, it can be assumed that a virus arrives at a random time between two activations of an edge. Therefore, the probability distribution of the

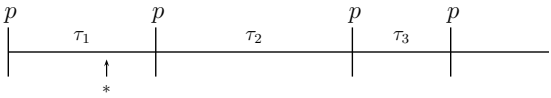

Figure 5.1 – Illustration of the process over an edge. The inter-event times τ_i are independent and randomly taken from the distribution $f(\tau)$. The initial infection of one end of the edge takes place at a random time with respect to the edge, illustrated by a star. At each activation of the edge, transmission of the virus will succeed with probability p.

success of the first attempt at time t follows the equation (1.5.21):

$$g(t) = \frac{1}{\langle\tau\rangle}\int_t^{+\infty} f(\tau)\,\mathrm{d}\tau, \tag{5.2.1}$$

where $\langle\tau\rangle$ is the mean inter-event time. Let us now turn to the case of an arbitrary value of p. From the previous expression, the inter-success distribution probability $\varphi(t)$ after an arbitrary number of attempts is given by

$$\varphi(t) = \sum_{k=0}^{\infty} p(1-p)^k P(k+1,t), \tag{5.2.2}$$

where $p(1-p)^k$ is the probability of a success after $k+1$ trials, and

$$P(k+1,t) = (g * f^{*k})(t), \tag{5.2.3}$$

where $*$ stands for the convolution product and $P(k,t)$ is the inter-success probability in k trials. This expression takes a simple form in the associated Laplace domain

$$\tilde{P}(k+1,s) = \tilde{g}(s)\tilde{f}^k(s), \tag{5.2.4}$$

where the upper tilde stands for the Laplace transform.

We will now determine the expectation $\langle t_s\rangle$ of the inter-success time distribution $\varphi(t')$ based on the moments of the inter-event times distribution $f(t)$.
Let us first recall the second-order small s expansion of each distribution:

$$\tilde{f}(s) = 1 - s\langle\tau\rangle + \frac{s^2}{2}\langle\tau^2\rangle + \mathcal{O}(s^3), \tag{5.2.5}$$

$$\tilde{g}(s) = 1 - s\langle t\rangle + \frac{s^2}{2}\langle t^2\rangle + \mathcal{O}(s^3). \tag{5.2.6}$$

Newton's generalized binomial theorem provides an expression for $\tilde{f}^k(s)$:

$$\left(\tilde{f}(s)\right)^k = (1 - s\langle\tau\rangle + \frac{s^2}{2}\langle\tau^2\rangle + \mathcal{O}(s^3))^k \tag{5.2.7}$$

$$= \sum_{j=0}^{k} \binom{k}{j} \Big(1 - s\langle\tau\rangle\Big)^{k-j} \left(\frac{s^2}{2}\langle\tau^2\rangle\right)^{j} + \mathcal{O}(s^3) \tag{5.2.8}$$

$$= (1 - s\langle\tau\rangle)^k + k\Big(1 - s\langle\tau\rangle\Big)^{k-1} \left(\frac{s^2}{2}\langle\tau^2\rangle\right) + \mathcal{O}(s^3) \tag{5.2.9}$$

$$= \sum_{j=0}^{k} \binom{k}{j} \Big(-s\langle\tau\rangle\Big)^{j} + k\frac{s^2}{2}\langle\tau^2\rangle + \mathcal{O}(s^3) \tag{5.2.10}$$

$$= 1 - ks\langle\tau\rangle + \frac{k(k-1)}{2}s^2\langle\tau\rangle^2 + k\frac{s^2}{2}\langle\tau^2\rangle + \mathcal{O}(s^3) \tag{5.2.11}$$

$$= 1 - ks\langle\tau\rangle + \frac{ks^2}{2}\Big((k-1)\langle\tau\rangle^2 + \langle\tau^2\rangle\Big) + \mathcal{O}(s^3). \tag{5.2.12}$$

Moreover, standard integrations of $g(t)$ allow to express its moments with respect to the ones of $f(\tau)$:

$$\tilde{g}(s) = 1 - s\frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + \frac{s^2}{2}\frac{\langle\tau^3\rangle}{3\langle\tau\rangle} + \mathcal{O}(s^3). \tag{5.2.13}$$

Incorporating equations (5.2.12) and (5.2.13) into equation (5.2.4) and keeping only the terms of degree one and two leads to

$$\tilde{g}(s)\tilde{f}^k(s) \approx \Big(1 - s\frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + \frac{s^2}{2}\frac{\langle\tau^3\rangle}{3\langle\tau\rangle}\Big) \times \Big(1 - ks\langle\tau\rangle + \frac{ks^2}{2}\Big((k-1)\langle\tau\rangle^2 + \langle\tau^2\rangle\Big)\Big) \tag{5.2.14}$$

$$\approx 1 - s\Big(k\langle\tau\rangle + \frac{\langle\tau^2\rangle}{2\langle\tau\rangle}\Big) + \frac{s^2}{2}\Big(k(k-1)\langle\tau\rangle^2 + 2k\langle\tau^2\rangle + \frac{\langle\tau^3\rangle}{3\langle\tau\rangle}\Big). \tag{5.2.15}$$

Finally, incorporating equation (5.2.15) into equation (5.2.2) allows us to determine the expectation $\langle t_s\rangle$ of the inter-success distribution $\varphi(t)$:

$$\langle t_s\rangle = \sum_{k=0}^{+\infty} p(1-p)^k \langle P(k+1,t)\rangle \tag{5.2.16}$$

$$= p\frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + \sum_{k=1}^{+\infty} p(1-p)^k\Big(k\langle\tau\rangle + \frac{\langle\tau^2\rangle}{2\langle\tau\rangle}\Big) \tag{5.2.17}$$

$$= p\frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + (1-p)\frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + \langle\tau\rangle\frac{1-p}{p} \tag{5.2.18}$$

$$= \frac{\langle\tau^2\rangle}{2\langle\tau\rangle} + \frac{1-p}{p}\langle\tau\rangle. \tag{5.2.19}$$

Note that we recover the expressions of the standard bus paradox when setting $p = 1$.

Decreasing p systematically increases the average inter-success time as expected, because more and more trials are required for the spreading to actually take place. In order to account for this trivial effect and to properly identify how the shape of the inter-event distribution affects $\langle t \rangle$ as a function of p, we focus on the normalized average success time $\bar{\tau}$, a standard measure for the burstiness of a process [Karsai et al. 2012; Kivelä et al. 2012], defined as

$$\bar{\tau} = \frac{\langle t_s \rangle}{\langle \tau \rangle / p}, \tag{5.2.20}$$

where $\langle \tau \rangle / p$ is the average success time in the case of a Poisson process. In the latter case, the inter-event time is exponential and $\langle \tau^2 \rangle = 2 \langle \tau \rangle^2$. One directly finds

$$\bar{\tau} \quad = \quad 1 + p \left(\frac{\langle \tau^2 \rangle}{2 \langle \tau \rangle^2} - 1 \right), \tag{5.2.21}$$

an expression taking the value 1 in the case of a Poisson process, as expected, and depending linearly on the burstiness coefficient β that we introduced in Section 1.4,

$$\beta \quad = \quad \frac{\langle \tau^2 \rangle}{2 \langle \tau \rangle^2}, \tag{5.2.22}$$

which takes a large positive value for heavy-tailed distributions, as the ones observed in a majority of empirical systems. This result clearly shows that burstiness tends to slow down the dynamics, but that its impact becomes less and less important as the probability of success p decreases and a higher number of attempts is necessary for the virus to spread.

5.2.2 Epidemic threshold and spreading efficiency

The dynamics described so far allows us to estimate the spreading of the infection in situations when nodes do not recover from the disease. In a majority of practical situations however, nodes remain infected only during a finite time before recovering. In that case, the propensity of the virus to invade the system is mainly governed by the connectivity of the network, e.g. the number of contacts per infected user, and its transmissibility $\mathbb{P}$, also called infectivity, defined as the probability that the virus will spread to an available neighbour before the infected node recovers. The transmissibility is defined as

$$\mathbb{P} \quad = \quad \int_0^{+\infty} \varphi(t) \int_t^{+\infty} r(\tau) \, \mathrm{d}\tau \, \mathrm{d}t, \tag{5.2.23}$$

where, as before, $\varphi(t)$ is the probability of a successful infection at time t, and $r(\tau)$ is the probability density for an infected node to recover at an ulterior time $\tau > t$. This expression clearly shows the importance of the competition between two temporal processes. In the case of tree-like networks, where all nodes have the same transmissibility $\mathbb{P}$, it is straightforward to show that the basic reproduction number, defined as

the average number of additional people a person infects before recovering, is given by

$$R_0 = \mathbb{P}\langle n \rangle, \tag{5.2.24}$$

where $\langle n \rangle$ is the expected number of susceptible neighbours of an infected node. The epidemic threshold is defined by the condition $R_0 = 1$ separating between growing and decreasing spreading. The epidemic threshold is thus reduced either by reducing the transmissibility or $\langle n \rangle$. This result is valid only when the network has a tree-like structure, which is valid for a majority of random network models below the epidemic threshold [Iribarren and Moro 2011].

In order to illustrate these results and to perform numerical tests in the following, we consider an inter-event distribution given by a Gamma distribution

$$f(t;a,b) = \frac{1}{b^a \Gamma(a)} t^{a-1} e^{-\frac{x}{b}}$$

with scale parameter b and shape parameter a. This family of distributions has the advantage of including the exponential distribution, for $a = 1$, and to provide distributions with a tuneable variance by changing a. One finds

$$\langle \tau \rangle = ab \tag{5.2.25}$$

$$\sigma^2 = \frac{\langle \tau \rangle^2}{a} \tag{5.2.26}$$

Fixing $\langle \tau \rangle$, one thus finds that the variance increases by decreasing a.

Let us also note that transmissibility takes a particularly simple expression when f is an exponential distribution $\lambda e^{-\lambda t}$. As there is a probability p of success of transmission at each attempt, the mean number of attempts of transmission before success is $\frac{1}{p}$. Since there is no bus paradox when f is exponential, $\varphi(t)$ is therefore an exponential with rate λp.

When the recovery time is a constant t_c, $r(t) = \delta(t - t_c)$, the transmissibility is given by

$$\mathbb{P} = \int_0^{+\infty} \lambda p e^{-\lambda p t} \int_t^{+\infty} \delta(t - t_c) \, \mathrm{d}\tau \, \mathrm{d}t \tag{5.2.27}$$

$$= \int_0^{t_c} e^{-\lambda p t} \, \mathrm{d}t \tag{5.2.28}$$

$$= 1 - e^{-\lambda p t_c}. \tag{5.2.29}$$

When the recovery distribution is an exponential with rate $\frac{1}{t_c}$, the transmissibility is given by

$$\mathbb{P} = \int_0^{+\infty} \lambda p e^{-\lambda p t} \int_t^{+\infty} \frac{1}{t_c} e^{-\frac{\tau}{t_c}} \, \mathrm{d}\tau \, \mathrm{d}t \tag{5.2.30}$$

$$= \int_0^{+\infty} \lambda p e^{-\lambda p t} e^{-\frac{t}{t_c}} \, \mathrm{d}t \tag{5.2.31}$$

$$= \frac{\lambda p t_c}{\lambda p t_c + 1} \tag{5.2.32}$$

As a first check, one shows the accuracy of equation (5.2.24) to determine the epidemic threshold on a tree by numerical simulations in Fig. 5.2.

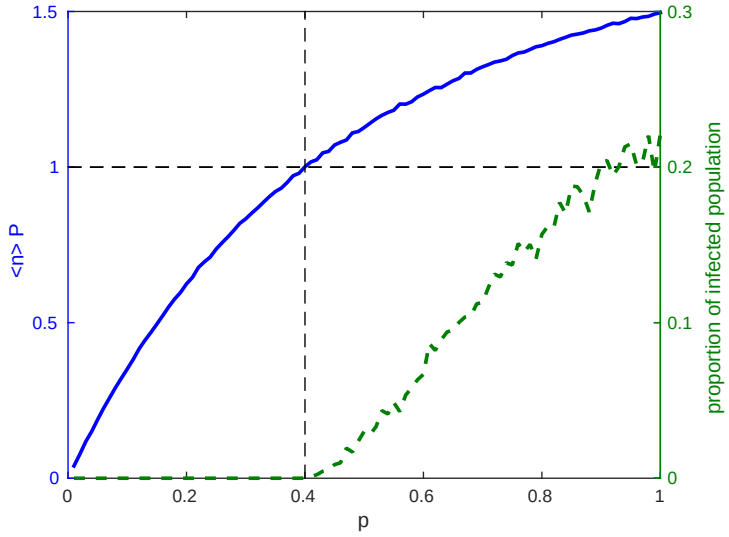

Figure 5.2 – Proportion of infected population (dashed green) and reproduction number (solid blue) for a tree-like network as a function of p. $\langle n \rangle \mathbb{P} = 1$ indicates the epidemic threshold. The parameters of the model are $\langle n \rangle = 2$, $r(t) = \delta(t-1)$, and the inter-event distribution is a Gamma distribution, with parameters $a = 0.5$ and $b = 1$.

Let us now consider the impact of the shape of the Gamma distribution, calibrated by a, on transmissibility, by using the exponential case $a = 1$ as a baseline for comparison. Numerical solutions show in Fig. 5.3 the dependency of transmissibility on $\langle \tau \rangle$ and p, for three values of a (the value of b is thus determined via equation (5.2.25)). Simulations are all performed with $r(t) = \delta(t-1)$ but similar results are obtained for exponential distributions for the recovery time. One observes that transmissibility depends on the ratio $\langle \tau \rangle / p$ when $a = 1$, but that the system exhibits deviations to this linear relationship when the dynamics deviates from a Poisson process. When $a \neq 1$, the isolines are indeed convex and cross the isolines corresponding to the Poisson case. In order to quantify this deviation, we explore the dependency of $\mathbb{P}$ on the success probability p, for fixed values of the parameter $\langle \tau \rangle / p$. The latter is simply a naive estimation of the time of success, obtained by multiplying the average time between two contacts $\langle \tau \rangle$ and the expected number $\frac{1}{p}$ of contacts before a success takes place.

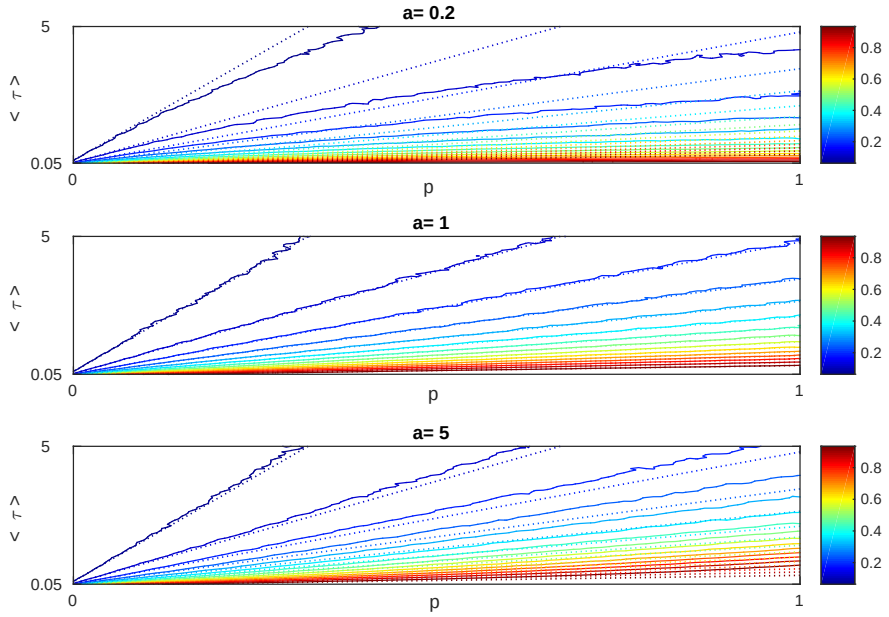

Figure 5.3 – Isolines for the transmissibility $\mathbb{P}$ in the $(\langle\tau\rangle, p)$ plane, for Gamma distributions with different shape parameters a and $r(t) = \delta(t-1)$. In the case of a Poisson process, isolines are linear, indicating a dependency of $\mathbb{P}$ on $\langle\tau\rangle/p$. Dashed lines indicate the linear relationship of a Poisson process. Deviations from $a = 1$ alter the convexity of the isolines.

Numerical results clearly show, in Fig. 5.4, that transmissibility becomes less and less efficient, as compared to the Poisson case, for increasing values of p, in agreement with the observation that the inter-success time is more and more affected by the bus paradox for increasing values of p. The mechanism behind this effect is also apparent in Fig. 5.5 where, again for a fixed value of $\langle\tau\rangle/p = 1$, transmissibility is clearly less efficient in the case of bursty dynamics for larger values of p. Intuitively, the reason is that, for large p, the first inter-event time, which is given by the bus paradox formula, plays a major role in determining the inter-success time because few attempts are required. On the contrary, when p is very low, many attempts are required before the success of the transmission. Therefore, the inter-success time is the sum of many random variables and follows a distribution that is obtained through the convolution product of many distributions (once g and many times f), which tends to "flatten" the resulting distribution. The decrease of the transmissibility $\mathbb{P}$ with respect to p when the variance of the inter-event distribution is larger may also be intuitively explained by the fact that few attempts are required when p is larger, giving more importance to large values of inter-event times, which are more likely to occur when the variance of the distribution is large.

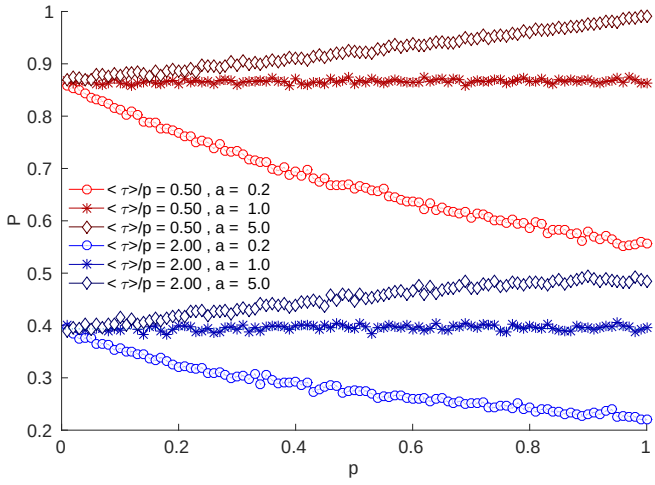

Figure 5.4 – Transmissibility $\mathbb{P}$ vs probability p of success for fixed values of $\langle\tau\rangle/p$, with $r(t) = \delta(t-1)$. In the case of Poisson process, with $a = 1$, the system does not exhibit a dependency on p. When the system is bursty, in contrast ($a < 1$), increasing p decreases the transmissibility of the process.

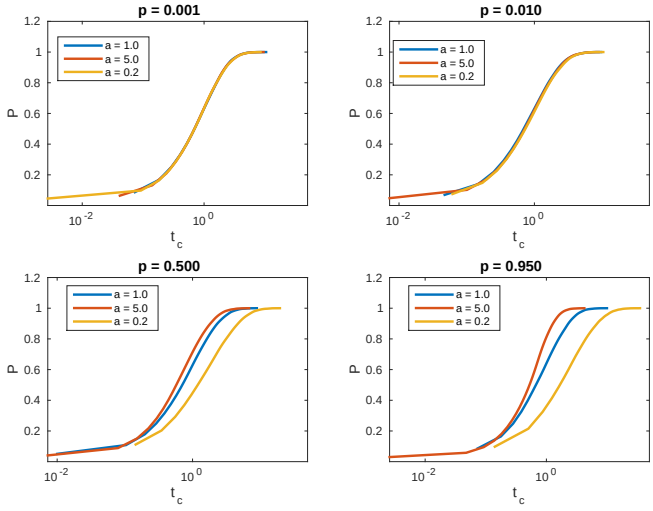

Figure 5.5 – Transmissibility as a function of the recovery time t_c, for $r(t) = \delta(t - t_c)$ and fixed values of $\langle\tau\rangle/p = 1$. The dynamics is more and more affected by the shape of the distribution for larger values of p.

This result has important practical implications, for instance in online marketing, where users' activity is known to be bursty. Indeed, let us consider a marketing agency having to decide on a strategy in order to optimize its viral impact. The impact of its campaign, evaluated by its transmissibility, increases only sub-linearly with its quality evaluated by its success probability p. This observation therefore needs to be considered when devising a strategy to balance the gains and the costs of increasing the viral potential of an ad.

5.3 Discussion

The main purpose of this chapter was to study the impact of bursty behaviour on the spreading on temporal networks when several attempts may be required for the virus to spread between two individuals. In contrast with a majority of previous works, we have considered a model where the transmission of the diffusive entity is only successful with a certain probability at each contact. Our main results are twofold. First, we have derived an analytical expression for the average inter-success time. We have shown that burstiness plays a more and more important role associated to a slowing down of the dynamics by the bus paradox when the probability of success is increased. In the limit of a small probability of success, the shape of the inter-event time distribution ceases to play a role, and only its average determines the speed of spreading. Second, we have turned to numerical computations in order to calculate the epidemic threshold for a dynamics where nodes remain infectious during a random period of time before recovering. Our results confirm that the bus paradox hinders the spreading of the process when the probability of success is increased. As we discussed, the results have implications for the design of efficient marketing campaigns. In particular, it suggests that an efficient strategy should aim at properly predicting the future activation times of a target user, in order to minimize the time between the infection and the first subsequent contact, and therefore the impact of the bus paradox. In this work, we have considered locally tree-like structure, as often assumed when studying spreading processes, especially in early phase. An extension of this work would be to incorporate more complex structure, where cycles and communities play a role, in particular to observe the impact of burstiness on the spreading around bottleneck users [Delvenne et al. 2015].

Chapter 6

Temporal sequence of retweets reveals cascade migration

This chapter presents the results of [Bhowmick et al. 2017].

Abstract

Twitter has recently become one of the most popular online social networking websites where users can share news and ideas through messages in the form of tweets. As a tweet gets retweeted from user to user, large cascades of information diffusion are formed over the global network. Existing works on cascades have mainly focused on predicting their popularity in terms of size. In this work, we leverage on the temporal pattern of retweets to model the spreading dynamics of a cascade. Notably, retweet cascades provide two complementary information: (i) the inter-retweet time intervals of retweets, and (ii) the spreading of the retweets over the underlying follower network. Using datasets from Twitter, we identify two types of cascades based on the presence or the absence of an early peak in their sequence of inter-retweet intervals. We identify multiple diffusion localities associated with a cascade as it propagates over the network. Our studies reveal that the transition of a cascade to a new locality is facilitated by *anchor nodes*, that are highly cascade dependent, following the saturation of the current locality. We propose an analytical model to explain co-occurrence of the first peak with the migration of the cascade to a new locality as well as to detect locality saturation from inter-retweet intervals. Finally, we validate these claims from empirical data showing a co-occurrence between the first peak and the migration with good accuracy; we obtain an even better accuracy for successfully classifying saturated and non-saturated diffusion localities from inter-retweet intervals.

6.1 Introduction

We will now specifically focus on information diffusion on online social media such as Twitter [Toriumi et al. 2013; Tonkin et al. 2012] through the means of real datasets. Note that information diffusion is actually a spreading process, however the standard terminology refers to it as information diffusion rather than information spreading. We will accordingly adopt this abuse of language in this monograph.

In Twitter, the spreading of tweets occurs via the retweeting of an initial tweet by the users. After each retweet event, a new set of users gets exposed to the content and popular tweets finally get widely retweeted, forming cascades that spread on the underlying social network.

Importantly, the spread of retweets on the underlying follower network is not always uniform. As we will show, the cascades may rather get spread in bursts from one locality of the network to another, caused by the *anchor users*, or *anchor nodes*. As a consequence, the corresponding audience reached by two cascades of the same size may substantially differ if one of them spread through different localities whereas the other remained confined within a single one [Weng et al. 2014].

The question whether the cascade spread in several localities, and to whom such a transition is due to, may be answered by exploiting the structure of the underlying social network. However, such information about the global structure is generally hard, costly -and sometimes impossible- to obtain as one requires the full list of contacts of each user. Our objective is therefore to explore the cascades' transitions only using the temporal information of the cascade, which is of interest as this kind of information is typically cheaper, easier and less intrusive to obtain than the structural ones.

In the previous chapters, we have shown that temporal characteristics of the spreading process may leave footprints on the topological observations of the spreading entity, for instance by inducing biases in the trajectory of a random walker. In a similar fashion, we will show in this chapter how the temporal patterns of a cascade, measured as inter (re)tweet intervals, may work as an indicator for cascade transition across diffusion localities. Inter-retweet intervals only related to the users, that is the time between two consecutive posts of a single user, have previously been exploited in the literature in order to classify users and detect bots [Tavares and Faisal 2013; Ghosh et al. 2011]. In this work, we focus on inter-retweet intervals related to the cascades rather than to the users.

We first present the Twitter datasets that we will use in both this chapter and the next one. Then, we empirically explore the time series of the cascade through the corresponding inter-retweet intervals, allowing to sort cascades into two types depending on the presence of a first peak in the inter-retweet intervals, and introduce the concepts of diffusion locality and its saturation in this context. Next, we propose a simple analytical model that explains the co-occurrence between the first peak and the migration of the cascade to a new diffusion locality. Afterwards, we validate the model with empirical observations where we confirm a co-occurrence between the first peak and the migration of the cascade, and show that the model successfully classifies the saturated and non-saturated diffusion localities from sequence of inter-retweet intervals with about 90% accuracy.

6.2 Datasets

In this work, we rely on two types of public Twitter datasets, depending on the information they contain. The first kind, that we refer to as *schematic* dataset, mostly contains a collection of tweets posted during famous events such as *15-M Movement* (anti-austerity movement in Spain) and *IPL 2018* (Indian Cricket Premier League) in addition to tweets posted by the pop-star *Lady Gaga*[(1)]. The second kind, that we call *comprehensive* dataset, are related to the Arab-Spring movement[(2)] (*Algeria and Egypt datasets*) [Bruns et al. 2013] and the *2015 Nepal earthquake*[(3)]. The variety of the topics of the different datasets (social movements, natural disaster, sport, pop-star celebrity) allows us to draw general conclusions that are not specific to a particular type of topics.

Both types of datasets contain the following information about the tweets and retweets, which are described in Table 6.1:

- The (re)tweet ID;
- The ID of the user who posted the (re)tweet;
- The timestamp of the (re)tweet;
- The ID of the original tweet (for the retweets), which allows to recreate the time series of a cascade.

Dataset	#Tweets	#Retweets	#Cascades	#Users	Maximum cascade size
Algeria	65268	17269	5730	1.6×10^4	980
Egypt	671417	188090	67539	7.6×10^6	432
Nepal	26424	521938	26424	2.9×10^5	23864
IPL 2018	3884	197210	3884	2.5×10^4	99
15-M	2626	5649951	2626	8.7×10^4	119424
Lady Gaga	1238135	6329596	1238135	1.6×10^5	14130

Table 6.1 – Details of the Twitter datasets.

Importantly, comprehensive datasets contain the additional information of follower/followee network of all the users who participated in a cascade from the dataset. Therefore, these comprehensive datasets have the fundamental advantage of providing information of (re)tweet spreading and the underlying social (follower) network, which makes them rich and expensive to collect.

(1) `http://www.cnergres.iitkgp.ac.in/blog/2019/03/08/twitter-cascade-dataset/`
(2) `http://www.cnergres.iitkgp.ac.in/blog/2018/02/28/arab-spring-twitter-dataset/`
(3) `http://crisisnlp.qcri.org/lrec2016/content/2015_nepal_eq.html`

In the following, we will mainly rely on the comprehensive *Egypt* and *Algeria* datasets that provide the underlying structure as ground-truth. Another advantage of these datasets is that the users live in the same timezone, which allows us to study the daily cycles. Our preliminary analysis on the comprehensive *Nepal* datasets provided results similar to the ones presented for the *Egypt* and *Algeria* datasets in the following. We will only use the other datasets to support the existence of two types of cascades based on the inter-retweet intervals on more datasets.

From the datasets, we recreate the sequence of retweets associated to a cascade. For a cascade C of size n_C originated by user u_0^C at time t_0^C, we have the time series of retweets ordered based on timestamps denoted by $(t_0^C, t_1^C, ,\ldots, t_{n_C}^C)$ with the corresponding list of retweeting users $(u_0^C, u_1^C, ,\ldots, u_{n_C}^C)$. Given this time series, we simply define the sequence of inter-retweet time intervals denoted by $T^C = (T_0^C, T_1^C, \ldots, T_{n_C-1}^C)$ for the cascade C as the time interval between two consecutive retweets in C such that the i^{th} inter-retweet time interval is computed as $T_i^C = t_{i+1}^C - t_i^C$. This means that an inter-retweet interval is the time between two consecutive retweets of the original tweet. Such retweets may occur in two different parts of the network. Therefore, this is not in general the time between the retweet of a user and a retweet of one of her follower.

The Twitter follower network can be represented as a directed graph $G = (U, E)$ where $U = \{u_1, u_2, \ldots, u_N\}$ is the set of users and $E = \{(u_i, u_j) : u_i, u_j \in U\}$ is the set of directed links from u_i to u_j denoting the who-follows-whom relationship between users. $F(u)$ denotes the set of followers of a user $u \in U$.

In this work, we preprocess the dataset by filtering out all small-sized cascades with a number of retweets smaller than 10.

6.3 Inter-retweet intervals and spreading dynamics

6.3.1 Sorting cascades based on inter-retweet intervals

The study of the inter-retweet intervals T^C of different cascades C reveals the appearance of *peaks* in T^C, which correspond to large inter-retweet intervals between two consecutive retweet events followed by small ones, indicating a slowdown of the spreading process of the cascade. These peaks may appear at the intermediate phase of the spreading process of a cascade (early peak) and also at the last few retweets towards the late phase of almost all the cascades (late peak). The time at which the peak is observed varies widely across different cascades, as well as their magnitude, therefore we can discard circadian or any other underlying periodic effect as cause of these peaks.

The interval T_i^C $(1 \leq i \leq n)$ is defined as a peak of the sequence T^C if $T_i^C > \mu_{T^C} + c_p \sigma_{T^C}$, where μ_{T^C} and σ_{T^C} denote the mean and standard deviation of T^C and c_p is a constant. This definition directly results from a simple outlier detection tech-

nique [Seo 2006] and allows to numerically detect peaks for a cascade C. The peak T_i^C is called early peak if $\frac{i}{n} \leq c_e$, where $c_e \in [0,1]$ is a constant. We set $c_p = 2$ and $c_e = 0.8$ in our experiments but we tested other values of c_p and c_e and obtained similar results. Importantly, there may be several peaks associated to a cascade but we will focus on the first one that occurred.

Depending on the phase of occurrence of the *first peak* in T^C, we classify cascades into the following two types:

- **Type I Cascades:** Characterized by the occurrence of its first peak at an intermediate phase of their inter-retweet time series T^C (early peak), much before the occurrence of its last retweets (a typical example is provided at Figure 6.1 where one early peak is observed, however several early peaks may be observed as well).
- **Type II Cascades:** Characterized by the absence of peaks at the intermediate phase of T^C (late peak). The first peak is observed when the last few retweet events take place towards the end of the cascade spreading (typical example at Figure 6.1).

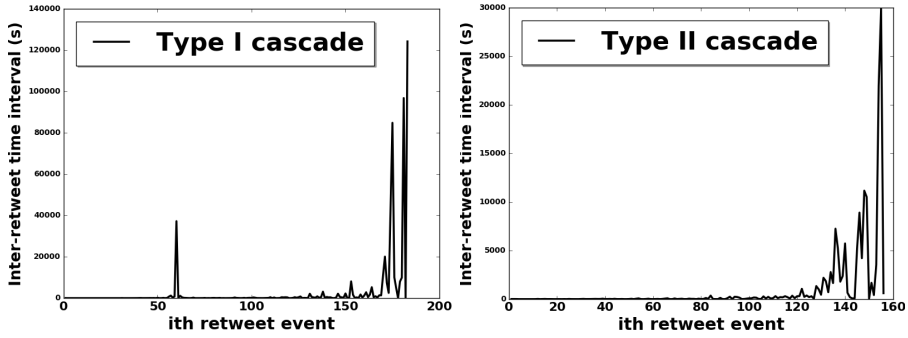

Figure 6.1 – Pattern of inter-retweet intervals for Type I and Type II cascades based on presence or absence of peaks in intermediate phase. Type I cascades admit at least one peak at an intermediate phase whereas all the peaks observed in Type II cascades occur around the last retweets.

Table 6.2 shows the fraction of Type I and Type II cascades observed across the six studied datasets. From this table, we observe that both types of cascades are detected in each dataset. We observe that datasets such as *Algeria*, *Egypt*, *15-M* and *IPL 2018* have a majority of cascades classified as Type I. On the other hand, the *Nepal* and *Lady Gaga* datasets have around 40% of Type I cascades. However, both these datasets consists in a large number of cascades implying that the number of Type I cascades for them is quite substantial even when the majority of the cascades are classified as Type II.

Dataset	Type I %	Type II %
Algeria	80	20
Egypt	61	39
Nepal	38	62
IPL 2018	58	42
Lady Gaga	37	63
15-M	56	44

Table 6.2 – Fraction of Type I and Type II cascades in the different datasets. Both types are detected in each dataset.

6.3.2 Cascade diffusion across diffusion localities

Let us now focus on the spreading of tweets over the underlying follower network. Let S^C denote the total set of exposed users for cascade C and S_i^C denote the set of exposed users after the first i retweets. We compute $P_i^C = \frac{|S_i^C|}{|S^C|}$ as the fraction of users in S^C who have been exposed to cascade C after the i^{th} retweet; thus P_i^C corresponds to the cumulative set of exposed users upto the i^{th} retweet. A typical plot of P_i^C after each retweet i ($1 \leq i \leq n$) is produced in Figure 6.2. Interestingly, one may observe that there is a sudden rise in P_i^C at some retweet i, or a *flush*, which indicates a sudden exposure of the content to a new population at the i^{th} retweet. Similarly to the definition of a peak, the i^{th} retweet ($1 \leq i \leq n$) is defined to cause a flush if $P_i^C > \mu_{P^C} + c_f \sigma_{P^C}$, where μ_{P^C} and σ_{P^C} denote the mean and standard deviation of P^C and c_f is a constant. We call *anchor user* the user who posted the retweet responsible of the flush. Otherwise, we denote *accretion* the rise in the exposed population when the tweet content gets newly exposed to a small population after retweet i.

We set $c_f = 3$ in our experiments but we tested other values of c_f and obtained similar results.

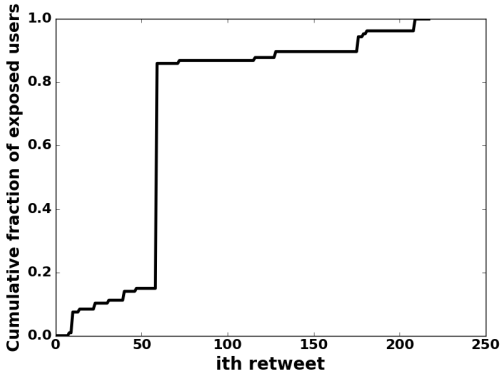

Figure 6.2 – Cumulative exposed population P^C for a Type I Cascade with flushes indicating exposure of the cascade to a new population.

In this context, we introduce the concept of diffusion locality L_a^C of a cascade C. We name it a diffusion locality rather than a spreading locality in order to echo with the "information diffusion" expression. Intuitively, a diffusion locality corresponds to a connected set of exposed users that incrementally grows with each retweet activity. After each retweet i, the locality of the current retweeter may incrementally grow by incorporating the set of newly exposed users. If the increase in the fraction of exposed users occurs due to an accretion, then the newly exposed population after retweet i will be absorbed within the current locality L_a^C. Otherwise, in case of a flush effect, this absorption process fails and a new locality L_b^C is discovered with retweet i, to which the cascade spreads. Thus, each new locality is discovered through an *anchor user* in the network, with a large non-exposed follower count. A flush in the sequence P^C causes the cascade C to spread to a new diffusion locality. Therefore, a diffusion locality associated to a cascade C is defined as the set of users who have been exposed to the content between two consecutive flushes in C, the first and the last retweet of C being considered as flushes by default.

6.3.3 Co-occurrence between early peaks and flushes

Interestingly, when superposing the plots of the inter-retweet intervals T^C and the fraction of exposed users P^C for a cascade C of Type I, one may observe that a flush in the exposed set of users occurs around the same time as the first peak, as typically illustrated in Figure 6.3. Only the first peak seems to co-occur with a flush. No flush typically co-occurs with the next peaks, and some flush may be observed long before the first peak. Importantly, such co-occurrence is not trivial because the only common information between T^C and P^C is the ordering of the tweets: T^C relies on purely temporal informations whereas P^C only relies on topological ones.

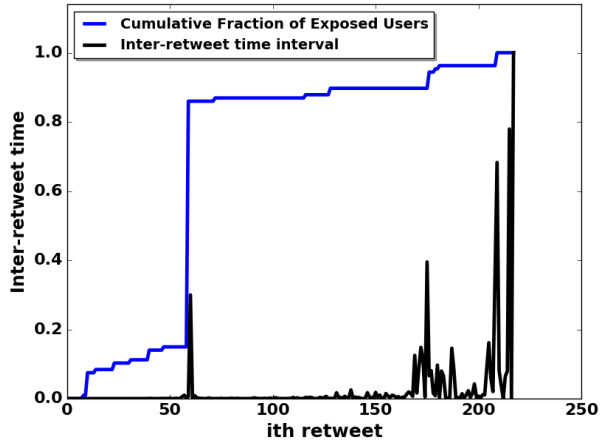

Figure 6.3 – The early peak in inter-retweet time intervals T^C of a Type I cascade typically co-occurs with a flush in the cumulative exposed population P^C.

We use the definitions of the flushes and of the peaks to detect them for each Type I and Type II cascade of the comprehensive datasets, and allow a maximal shift of three events to consider that there is a co-occurrence. The results are shown in Figure 6.4 for Algeria and Egypt datasets. For the Type I cascades, we observe the co-occurrence between the first peak and a flush for 82% of cascades in Algeria dataset, and for 71% of such cascades in Egypt dataset. On the contrary, there is no co-occurence between flushes and peaks for Type II cascades.

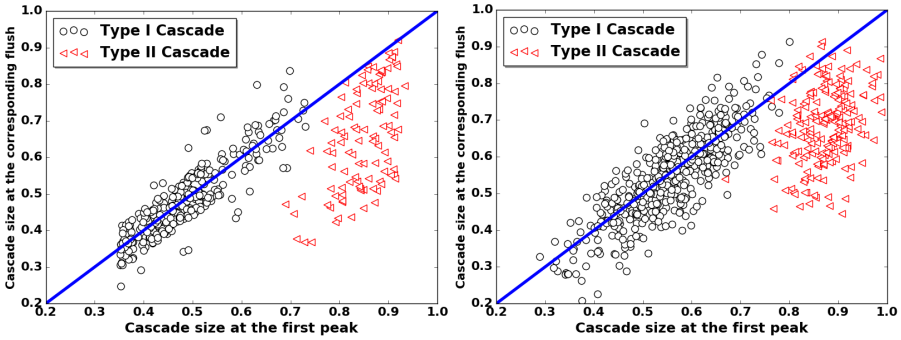

Figure 6.4 – Co-occurrence between a flush and the first peak appears for Type I cascades (blue circles) but not for Type II cascades (red triangles) in Algeria [Left] and Egypt [Right] datasets.

Therefore, it seems that detecting an early peak in Type I cascade allows to detect the exposure of the tweet to a new audience. This means that from the temporal sequence of retweet, one may draw conclusion about how the tweet propagated on the network, without requiring any knowledge of the underlying structure.

6.4 Analytical model for cascade spreading across diffusion localities

We now propose an analytical model that formalizes the process described before, and that allows us to explain the two empirical observations we have made:

1. The presence of a typical early peak in inter-retweet intervals for cascades of Type I and the absence of such a peak before last phase for cascades of Type II;

2. The co-occurrence between an early peak in inter-retweet intervals and the exposition of the tweet to a new diffusion locality for cascades of Type I.

6.4.1 Modeling cascade spreading in a single diffusion locality

The model considers the posting of a tweet from a seed node. As retweets take place, the initial diffusion locality of the seed becomes saturated, until a retweet reaches a new locality. Our derivations focus on the time evolution of the inter-retweet intervals generated by this process. A locality of size ν is approximated by a central node, the seed, with ν neighbors.

Let X_i be the retweet time instance of a user i and f be the corresponding probability distribution between two consecutive retweets of a user. Here, we assume that the time of retweets of the neighbors of the seed node are all conditionally independent and drawn from the same distribution f.

The sorted sample $(X_{(1)},\ldots,X_{(\nu)})$ is the observed time series of the cascade. The k^{th} smallest value $X_{(k)}$ is observed at time t with density $f_{(k)}(t)$, provided that exactly $k-1$ events occur before t (with probability $(F(t))^{k-1}$), and the $\nu-k$ remaining events occur after t (with probability $(1-F(t))^{\nu-k}$). Thus, $X_{(k)}$ follows the distribution $f_{(k)}$ given by

$$f_{(k)}(t) \quad = \quad \frac{\nu!}{(\nu-k)!\,(k-1)!}\, f(t)\, F(t)^{k-1}(1-F(t))^{\nu-k}, \tag{6.4.1}$$

where $F(t) \equiv \int_0^t f(\tau)\,\mathrm{d}\tau$ is the cumulative function of f.

Let E_k be the k^{th} inter-retweet interval, defined as $X_{(k+1)} - X_{(k)}$. The expectation of E_k denoted by $\langle E_k\rangle$ is given by:

$$\begin{aligned}
\langle E_k\rangle &= \int_0^{+\infty} t\left(f_{(k+1)}(t) - f_{(k)}(t)\right)\mathrm{d}t & (6.4.2)\\
&= \int_0^{+\infty} t f(t)\left[\frac{\nu!}{(\nu-k-1)!\,k!}\, F(t)^k(1-F(t))^{\nu-k-1}\right.\\
&\qquad \left. - \frac{\nu!}{(\nu-k)!\,(k-1)!}\, F(t)^{k-1}(1-F(t))^{\nu-k}\right]\mathrm{d}t. & (6.4.3)\\
&= \int_0^{+\infty} t f(t)\frac{\nu!}{(\nu-k)!\,(k)!}\\
&\qquad \times F(t)^{k-1}\,(1-F(t))^{\nu-k-1}\,\nu\left[F(t) - \frac{k}{\nu}\right]\mathrm{d}t & (6.4.4)
\end{aligned}$$

Let us first determine $\langle E_k\rangle$ considering a Markovian process, where $f(t) = \lambda e^{-\lambda t}$. In this case, an alternative approach, other than directly solving equation (6.4.4), allows us to determine $\langle E_k\rangle$ more easily.

Let us assume that the k^{th} smallest events occurred. Since the exponential distribution is memoryless, the time $\langle E_k \rangle$ we have to wait for the $(k+1)^{\text{th}}$ event to occur does not depend on how long we have already been waiting for the previous one. Therefore, it is fully determined by the mean of the distribution of the minimum of the $n-k$ events that will still occur, which is also an exponential with parameter $(n-k)\lambda$, which leads to

$$\langle E_k \rangle \quad = \quad \frac{1}{\lambda} \, \frac{1}{n-k} \tag{6.4.5}$$

which is a monotonically increasing function of k. Importantly, when the diffusion locality is not saturated ($k << \nu$), we obtain $\langle E_k \rangle \approx 0$, indicating the very low inter-retweet time intervals. On the other hand, when the locality approaches saturation ($k \approx O(\nu)$), we find $\langle E_k \rangle \approx 1/\lambda$, increasing the inter-retweet time intervals.

In situations where f takes a general form, corresponding to a general renewal process, the equation (6.4.4) cannot be explicitly solved. In the non-saturation phase when almost every member in the locality has not retweeted the message ($k << O(\nu)$), one may use the Stirling approximation of the factorial $\nu! \approx \sqrt{2\pi}\ \nu^{\nu+\frac{1}{2}}\ e^{-\nu}$ into equation (6.4.4) to obtain:

$$\langle E_k \rangle \quad \approx \quad \int_0^{+\infty} t f(t) F(t)^{k-1} \underbrace{\nu(\nu-k)^k (1-F(t))^{\nu-k-1}}_{\downarrow\ 0} \times \underbrace{\left(\frac{\nu}{\nu-k}\right)^{0.5}}_{\downarrow\ 1} \underbrace{\left(\frac{\nu}{\nu-k}\right)^{\nu}}_{\downarrow\ e^k} e^{-k}\, \mathrm{d}t$$

The integral tends to 0 when n tends to infinity, meaning that when there is no saturation, the inter-times remain low. Thus, $\langle E_k \rangle$ tends to be small when k decreases, implying larger values of $\langle E_k \rangle$ for the last inter-retweet intervals (high k) of a cascade due to the saturation.

This finding is confirmed by numerical simulations in Figure 6.5 for a variety of Gamma distributions, which also indicate that the values of the last inter-retweet times increase with the variance, while the behaviour in the non-saturation phase may vary depending on the system parameters. Therefore, the emergence of such a phenomenon is favoured when f is heavy-tailed, which is a typical observation in human-related systems.

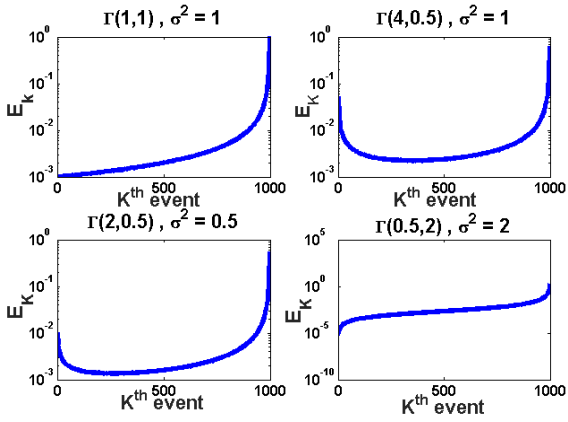

Figure 6.5 – Evolution of inter-retweet times when activity distribution is $\Gamma(a,b)$; we observe a high increase of the last inter-retweet times for all distributions at the saturation phase. Mean values over 50000 samples.

6.4.2 Modeling cascade migration across multiple localities

We now consider the interaction between two diffusion localities L_1 and L_2, which we model as two central nodes H_1 and H_2 respectively connected to each other and to their neighbors. Again, we assume that the instances of retweets of all the users are all conditionally independent and drawn from the distribution f.

The first central node H_1 triggers a post which gets exposed to its ν_1 neighbors, including H_2, generating retweet instances $(X_1,\ldots,X_{\nu_1})$. Potentially, when H_2 retweets the post, it leads to an exposure to the new locality formed by H_2 and its ν_2 neighbors, generating retweet instances $(Y_1,\ldots,Y_{\nu_2})$. In the observed (sorted) time series of the cascade $(Z_{(1)},\ldots,Z_{(\nu_1+\nu_2)})$, let the p^{th} retweet correspond to the retweet of H_2, and the q^{th} to the one of the first neighbor of H_2 who retweets ($q > p$).

Similar to E_k, we introduce F_k as the k^{th} inter-retweet interval of the cascade, defined as $Z_{(k+1)} - Z_{(k)}$. From the definition of p and q, we find $Z_{(k)} = X_{(k)}$ for $k < q$ (spreading within L_1), $Z_{(q)} = X_{(p)} + Y_{(1)}$ (first retweet in L_2, locality of H_2) and $Z_{(q+1)} = \min(X_{(p)} + Y_2, X_{(q)})$ (subsequent retweets, in the locality of L_2 or L_1). Hence $F_k = E_k$ for $k < q$, and $F_{q+1} \leq Y_{(2)} - Y_{(1)}$, which attains a low value.
Now there can be two possibilities.

If locality L_1 saturates before retweets occur in L_2 ($q \approx O(\nu_1)$), F_{q-1} observes a rise (since $F_{q-1} = E_{q-1}$ and following section 6.4.1, E_{q-1} gets a high value), and subsequently $F_{q-1} >> F_{q+1}$. This results in a sudden rise and a sharp fall in the consecutive values of F_k, showing a peak in the time series at events $q-1$ to q, and the corresponding cascade is classified as a Type I cascade. Since q is lower bounded by p, this pattern will also be observed when H_2 retweets after L_1 saturates ($p \approx O(\nu_1)$).

On the contrary, if locality L_2 retweets before L_1 gets saturated ($q << O(\nu_1)$), F_{q-1} will get a low value (following section 6.4.1) and thus the decrease in F_k will be

too small for the fall to be observed as both F_{q-1} and F_{q+1} are low. The corresponding cascade will therefore be classified as a Type II cascade.

Depending on the presence of such a second locality L_2 and the value of p, we observe the following three cases, illustrated through numerical simulations in Figure 6.6.

(a) **Case 1:** Cascade starts propagating from H_1 and reaches saturation observed by a peak in E_k; H_2 does not retweet, therefore the cascade does not reach its associated locality L_2. This situation produces a Type II cascade.

(b) **Case 2:** Cascade propagates from H_1; H_2 retweets and cascade reaches L_2 at an early phase, when L_1 is not yet saturated ($p << O(\nu_1)$). The lack of timescale separation for retweets in L_1 and L_2 does not allow for the observation of a peak. This situation produces a Type II cascade.

(c) **Case 3:** Cascade starts from H_1; H_2 retweets and cascade reaches L_2 at a late phase, when locality L_1 is already saturated ($p \approx O(\nu_2)$). The consecutive retweet events happen around H_2, and thus small values of the inter-retweet times in H_2 follow the large values in H_1 which leads to an observation of a peak at an intermediate stage. This situation produces a Type I cascade.

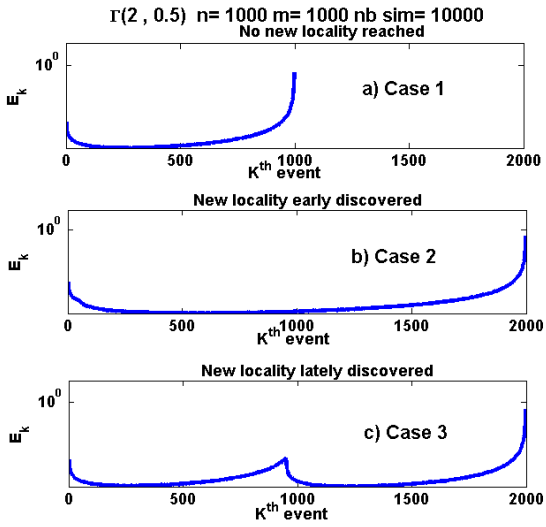

Figure 6.6 – Migration to a new locality when inter-retweet times follow $\Gamma(a,b)$ distribution. Cascades of Type II are observed when there is no migration or when migration to L_2 happens before saturation of L_1 (case 1 and 2); cascades of Type I (early peak) are observed for migration to L_2 after saturation of L_1 (case 3). Mean values over 50000 samples.

The global dynamics behind the formation of Type I and Type II cascades is schematically summarized in Figures 6.7 and 6.8.

In a nutshell, our model demonstrates that inter-retweet intervals E_k provide signatures of the content saturation in the current diffusion locality, and also that the presence of an early peak in the inter-retweet interval (manifested by the rise and fall in E_k) indicates the migration of the cascade to a new diffusion locality after saturation of the current locality.

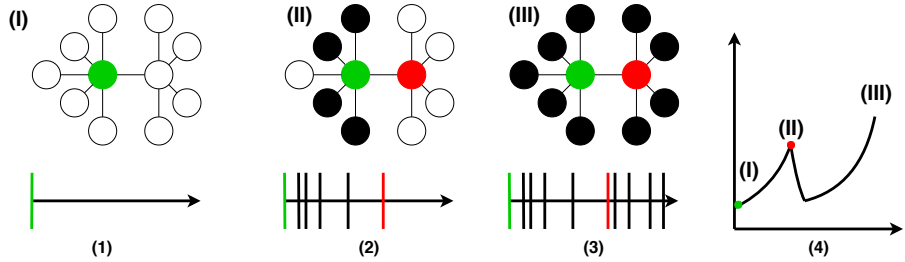

Figure 6.7 – Dynamics of Type I cascades. Below each phase of the cascade spreading process, the time series of the cascade is displayed, where each stroke on the timeline corresponds to a retweet. The corresponding sequence of inter-retweet intervals is displayed in (4). First, the seed H_1 tweets (I, in green). When the tweet saturates the locality L_1 (II, retweeting nodes in black), the inter-retweet intervals increase (II). When it gets exposed to the new locality L_2 through the anchor node H_2 (in red), a new wave of spreading occurs in the second locality L_2, resulting in a fall of the inter-retweet intervals (III).

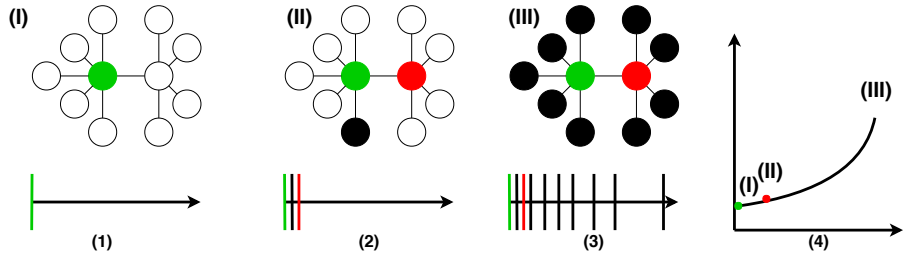

Figure 6.8 – Dynamics of Type II cascades. Below each phase of the cascade spreading process, the time series of the cascade is displayed, where each stroke on the timeline corresponds to a retweet. First, the seed H_1 tweets (I, in green). When the tweet gets exposed to a new locality L_2 through the anchor node H_2 (in red) before saturating the current locality L_1 , the inter-retweet intervals remain low (II). When most retweet events have occurred from users in both localities L_1 and L_2 (retweeting nodes shown in black), it results in a rise of the inter-retweet intervals (III).

6.5 Empirical validation

In this last section, we empirically validate the claims made in the analytical model. In order to explain the existence of Type I and Type II cascades and the co-occurrence of the first peak and a flush for Type I cascades, we developed a model that makes further assumptions by relating the inter-retweet intervals to the content saturation and migration in diffusion localities, which we will now investigate using the *Egypt* and *Algeria* datasets.

6.5.1 Cascade migration after the first peak

The analytical model assumes that retweets of Type I cascades mainly occur in the new diffusion locality after the first peak. In order to validate this model, we show that after the peak occurring at the p^{th} retweet, the retweeters mainly belong to a new diffusion locality.

We define four sets of users for the retweet sequence T^C of a Type I cascade C: (1) the set of users $\mathscr{P}$ retweeting around the first peak of inter-retweet intervals (therefore including the anchor node), (2) the set $F_{\mathscr{P}}$ of the followers of $\mathscr{P}$, (3) the set of users $\mathscr{B}$ who retweet before the p^{th} retweet (so before the users in $\mathscr{P}$) and their followers, and (4) the set of users $\mathscr{A}$ retweeting after p^{th} retweet (so after the users in $\mathscr{P}$).

From these sets, we compute the fraction of exposed users $f^C_{\mathscr{B}}$ after the first peak of cascade C, who already got exposed to the post prior to the peak (i.e., fraction of users in $\mathscr{A}$ that are also in $\mathscr{B}$); whereas $1 - f^C_{\mathscr{B}}$ denotes the fraction newly exposed to the tweet after the peak.

We compute the ratio $r^p_{new} = \frac{1-f^C_{\mathscr{B}}}{f^C_{\mathscr{B}}}$, as illustrated in Figure 6.9, that indicates whether the tweet mainly propagated further into a new locality. If $r^p_{new} > 1$, this indicates that the tweet had saturated its current locality and migrated to a new diffusion locality where it started a new wave of propagation.

In Figure 6.10, we illustrate the ratio r^p_{new} for all retweets i of a typical Type I cascade C. Importantly, we observe that this ratio becomes greater than 1 only for retweets when a peak occurs in sequence of inter-retweet intervals T_C.

More generally, we compute this ratio r^p_{new} for all the cascades. For 72% and 81% of Type I cascades in *Algeria* and *Egypt* dataset respectively, we obtain values of r^p_{new} larger than one always corresponding to some users retweeting near the peak. On the contrary this ratio r^p_{new} is always lower than 1 for cascades of Type II, indicating the absence of migration of the post from one locality to another.

This results seem to confirm that, for a majority of Type I cascades, the existence of an early peak in the inter-retweet intervals indicates the saturation of the initial diffusion locality and the subsequent migration of the cascade to a new locality.

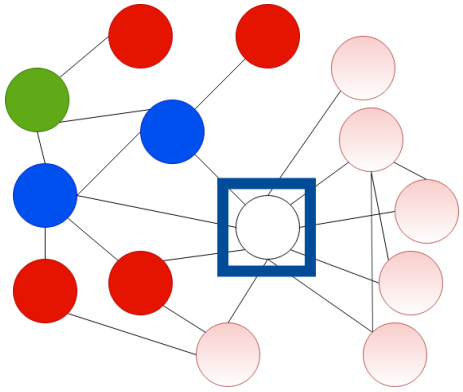

Figure 6.9 – Illustration of the four sets used to compute r^p_{new}. We assume that the seed (in green) and $p-1$ users have already retweeted (in blue). When the p^{th} user (pink squared in blue) retweet, r^p_{new} computes the ratio of newly exposed users (in pink) over the already exposed users who have not already retweeted (in red).

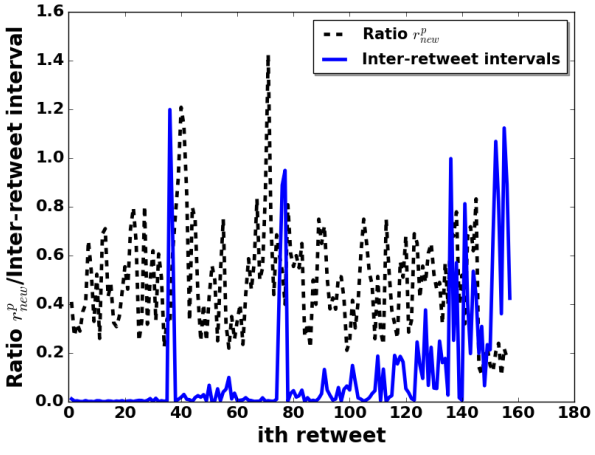

Figure 6.10 – Relation between first peak and flush effect indicating a new locality. We observe that r^p_{new} is larger than 1 for few users, always around a peak.

6.5.2 Locality saturation and inter-retweet intervals

Finally, we empirically validate the claim made in the analytical model that the inter-retweet time intervals are low ($E_k \approx 0$) when the diffusion locality is not saturated ($k << \nu$), and high ($E_k >> 0$) as it approaches saturation ($k \approx O(\nu)$). For each cascade of the datasets, we consider the inter-retweet interval at the first peak as the reference point; less than 20% of this interval can be considered as 'low' and more than 80% as 'high'. Notably, we repeated the experiments with a variety of thresholds; nevertheless, the outcomes remained consistent. In *Egypt* (resp. *Algeria*) dataset, we observe that when the saturation level ($\frac{k}{\nu}$) is lower than 0.3, 94% (resp. 92%) of corresponding inter-retweet intervals exhibit a low value. On the contrary, when the saturation level is greater than 0.8, 89% (82% for Algeria) of the corresponding inter-retweet intervals exhibit a high value.

Finally, the appearance of the majority (78% and 83% for *Algeria* and *Egypt* respectively) of the high inter-retweet intervals during the last few retweet events in all the cascades can be explained by the high variance of the user's activity distribution, which is typically heavy-tailed.

6.5.3 Remarks on anchor nodes

Anchor nodes play a key role in the flush effect which results in cascade migration across multiple diffusion localities. In principle, these users should possess a large number of followers that have not yet been exposed to the tweet content. This apparently makes the characterization of anchor nodes straightforward: the users with high follower count (or hubs) should be featured as potential anchor nodes. However, there exists an important distinction between hubs and anchor nodes which makes their identification challenging: the hub nodes can be characterized as the high degree nodes in the follower network. Thus it relies on the static structural property of the network. On the contrary, anchor nodes are cascade specific and depend on dynamics taking place in the cascade.

In *Algeria* and *Egypt* datasets, we observe that only 35% and 30% of the hubs act as anchor nodes. Moreover, we observe that only 38% and 42% of the flushes are caused by hubs, meaning that a significant amount of anchor nodes are lower degree nodes who succeeded in exposing the tweet for the first time to a large majority of their followers, contributing to the exposed population of the cascade. Therefore, the anchor nodes rely on other structural and temporal properties than just the degree.

The identification of such anchor nodes, which act as influential users for the information diffusion, will be the interest of the next chapter. In particular, we will show how to exploit the findings of this chapter in order to detect influential nodes when the structure of the underlying is unknown or only partially available.

6.6 Discussion

The main purpose of this chapter was to highlight how the temporal patterns of a cascade may be exploited to learn how it propagated on the underlying network. Our main result consisted in showing that the time sequence of retweet cascades carries a unique signature to detect the saturation of a tweet content within a single diffusion locality and its migration across multiple localities.

We have classified Type I and Type II cascades based on the presence or absence of an early peak in the inter-retweet time intervals, obtained from the retweet sequence. We have introduced the concept of diffusion locality associated to a cascade, which describes the population exposed to the content of a tweet. The tweet starts spreading inside an initial diffusion locality and may propagate to other diffusion localities through flushes in the exposed set of users caused by anchor nodes. We have proposed and empirically validated an analytical model which explains the co-occurrence of the first peak in inter-retweet intervals with a flush in the exposed set of users through the saturation and migration of cascades across new diffusion localities.

In this work, we have shown the existence of two types of cascades for several datasets. However, empirical validations on larger datasets would help validating the model. Our model successfully explains the existence of about 80% of the Type I cascades, an extension of this work would consist in focusing on the undestanding of the mechanism behind the formation of the remaining unexplaind 20%, for which a larger sample dataset is required.

Even though our method is data-driven and is limited by the fine-tuning of several thresholds, we have highlighted the key principle that the temporal patterns of the cascades reveal their structural exploration of the network. On the one hand, our work allows one to gain insights about whether a specific cascade propagated by bursts on the underlying topology. On the other hand, our work also allows one to detect the users who are responsible for the transition of the cascade to a new part of the network when the current part of the network where the tweet was propagating was saturated. Hence, having a large collection of cascades that propagated on a given network allows to detect the users who tend to act as an anchor node. Such users would therefore act as influential nodes for the information diffusion, and consequently may be preferentially targeted by smart spreaders. This idea of exploiting the history of cascades in order to detect influential nodes will be developed in chapter 7 .

Chapter 7

SmartInf: exploiting temporal sequences of retweets to detect influential users

This chapter presents the results of [Bhowmick et al. 2019].

Abstract

The identification of influential users in online social networks allows one to facilitate an efficient information diffusion to a large part of the network, thus being a great advantage to diverse applications including viral marketing, disease control and news dissemination. The existing methods have mainly only relied on the network structure for the detection of influential users. In this chapter, we enrich this approach by proposing a fast and efficient algorithm called *SmartInf* to detect a set of influential users by identifying *anchor nodes* from the temporal sequence of retweets in Twitter cascades. Such *anchor nodes* provide important signatures of tweet spreading across multiple diffusion localities and hence act as precursors for the detection of influential nodes. The set of influential nodes identified by *SmartInf* have the capacity to expose the tweet to a large and diverse population when targeted as seeds thereby maximizing the influence spread. Experimental evaluations on empirical datasets from Twitter show the superiority of *SmartInf* over the state-of-the-art baselines in terms of infecting a larger population; further, our evaluation shows that *SmartInf* is scalable to large-scale networks and is robust to missing data. Finally, we investigate the key factors behind the improved performance of *SmartInf* by testing our algorithm on a synthetic network using synthetic cascades simulated on it. Our results reveal the effectiveness of *SmartInf* in identifying a diverse set of influential users that facilitate a faster spreading of tweets to a larger population.

7.1 Introduction

We will now directly leverage on our findings of the chapter 6 to address the problem of multi-target selection in information diffusion.

In recent years, the research community has increasingly focused on whether spreading on online social networks can be maximized by sending a piece of information to certain special individuals, often called *influential users* [Stai et al. 2018; Dong et al. 2018]. These users may play a significant role in the spreading process, favouring a larger spread of the information, due to their activity as well as their position in the network. Once identified, these users, who typically represent a very small fraction of the network, may be targeted directly for efficient information spread. Existing literature on identifying influential users in a social network have mainly concentrated on using the knowledge of the underlying network topology [Huang et al. 2014; Xia et al. 2016; Madotto and Liu 2016; Goldenberg et al. 2018]. A brief review of the state-of-the-art on detecting influential users has been provided in section 1.6.3. Several works also focused on developing heuristics for the optimization of the set of seeds for the standard SI model, considering fixed [He and Kempe 2016; Tang et al. 2017; Wilder et al. 2017] or varying [Zhang et al. 2014] size of the seed set. Such optimization problem has been shown to be NP-hard [Kempe et al. 2003].

However, to the best of our knowledge, the state-of-the-art literature mostly overlooked the broad heterogeneity of users in terms of their temporal patterns, while identifying the influential nodes. Yet a user occupying a central position in the network, according to a given topological centrality measure, may not be a highly influential node despite this favorable structural position. Indeed, such a user may remain passive in posting the retweets, mostly participate in non-popular cascades, or retweet at the late phase of a popular cascade without helping it to spread further. On the other hand, overall activity rate, even for a structurally central user, does not reflect the true role in spreading tweets in popular cascades. Thus, the methods relying purely on the network structure and mean user activity are blind to these aspects, thereby may fail in their purpose by highly ranking some ineffective users who are very unlikely to reinforce the spreading process.

Moreover, state-of-the-art methods [Cha et al. 2010; Chen et al. 2012; Huang and Yu 2017] relying on network topology need to directly compute the influence of each node to identify and rank the influential users; this requires complete information of the network topology, which is typically difficult and expensive to obtain for large-scale social networks as we already mentioned. Additionally, utmost care needs to be taken when recommending a set of influential nodes to target simultaneously, as aiming influential nodes with overlapping followers would result in exposing twice the same audience to the content. As a side note, let us mention that re-exposure still helps the tweet to propagate [Bao et al. 2013; Shen et al. 2014; Gueuning et al. 2015], but exposing once the content to several parts of the network intuitively results in a larger spreading than exposing it several times to one single part of the network only.

In chapter 6, we have shown that exploiting the temporal sequence of retweets allows one to detect the anchor nodes, who play a key role in spreading the tweet from one diffusion locality to another. The objective of this work is to propose an algorithm leveraging on both the temporal retweet sequence of cascades and the local structural information to identify the set of influential users in the network. This approach has the benefits of not requiring the knowledge of the full topological structure of the network on one hand while taking into account the heterogeneous activity among users in a non-trivial way on the other hand.

We propose a scalable unsupervised algorithm *SmartInf*, that stands for *Smart Influential*, to detect the influential nodes in a social network. The proposed methodology relies on the peaks observed from the inter-retweet intervals of the cascades and on a refining step exploiting the follower list of some specific users.

This second step relies on the underlying topology of the network, but only some local structural information is required, as opposite to standard methods which require the full global structural information. If no information about the structure is available, we only exploit the history of cascades and name *SmartInf-Temp* this intermediate algorithm.

In this chapter, we first present the *SmartInf* algorithm and discuss its computational cost, the importance of its refining step and the standard centrality scores of the users of the ranked list it provides. Then we compare the algorithm to competing baseline algorithms for identifying the influential users in a social network. For this purpose, we use the comprehensive datasets introduced in chapter 6, and show that *SmartInf* outperforms the state-of-the-art baseline algorithms based on both Twitter-specific metrics as well as on numerical simulations. Following, we discuss the quality of the influential nodes obtained by *Smartinf* as well as its robustness. Finally, we turn to synthetic data to highlight some key factors behind the *Smartinf* performance.

7.2 The SmartInf algorithm

7.2.1 Problem statement

In this work, our objective is to determine a ranked list of influential users $\mathscr{S}$ in a social network G which can spread the information to a large population in G. First we assume that only the temporal retweet sequences T^C of a collection of cascades as well as the corresponding lists of retweeting users $(u_0^C, u_1^C, , \ldots, u_{n_C}^C)$ are available. Then we assume that it is possible to extract the follower list of some specific users from the global follower network G.

The goal is to identify a ranked seed set $\mathscr{S}$ of size β such that the mean exposed population $V(\mathscr{S})$ is maximized at the end of the spreading process, when $\mathscr{S}$ is targeted as a seed set. Mathematically, our problem can be formulated as identifying the ranked influencer set $\mathscr{S}$ such that for top-k influencers $\mathscr{S}^k \subseteq \mathscr{S}$, we obtain $max(V(\mathscr{S}^k))$.

7.2.2 SmartInf algorithm description

We hereby describe the SmartInf algorithm. The methodology consists in two major phases. The first phase (steps 1 and 2) consists in constructing a list $\mathscr{T}$ of influential nodes from the sequences of inter-retweet intervals T^C. The second phase (step 3) constructs the final ranked influencer list $\mathscr{S}$ from $\mathscr{T}$ using the followers' lists of the users in $\mathscr{T}$. We call *SmartInf-Temp* the truncated algorithm consisting in only applying the first phase. Thus, *SmartInf-Temp* provides a list of influential nodes only base on the temporal retweet sequence T^C of Type I cascades, which may be useful when the list of followers of the users cannot be obtained.

Step 1: Detection of potential influencers.

We first apply the peaks detection method described in section 6.3.1 to sort the cascades, and keep the set $\mathbf{T}$ of Type I cascades. For each cascade $C \in \mathbf{T}$, we know that the user u_p^C retweeted just before the first peaks, and that retweet of user u_q^C caused the consecutive fall in T^C. Since the first peak is usually accompanied by a flush in the set of exposed users, we claim that the anchor node is a user who retweeted between the p^{th} and the q^{th} tweet. Thus, the users retweeting within these tweets are denoted as the potential influential nodes I^C. Considering all the Type I cascades $\mathbf{T}$ in the dataset, we obtain the set of potential influential nodes $\sqcup$ of size α as $\sqcup = \bigcup_{C \in \mathbf{T}} I^C$. While constructing the set $\sqcup$, we keep track of the frequency of appearance of each user $u \in \sqcup$ in I^C.

Step 2: Ranked list of influencers $\mathscr{T}$.

The idea behind the step 1 was to identify the set of users I^C responsible for the migration of a Type I cascade C to a new diffusion locality. However, few sporadic users, who do not play any role in the migration of the cascade, may also appear in I^C and therefore get unfairly favoured. Nevertheless, the presence of such sporadic users in I^C is quite irregular across different cascades C; on the contrary, the truly influential users are likely to be consistently present in I^C since they retweet more often around the peak of Type I cascades. Hence, in Step 2, we rank the set of users $\sqcup$ according to their frequency of appearance in I^C; as a result, truly influential users should be ranked high. This procedure filters out the unfairly favoured users, who sporadically appear in I^C, placing them at the bottom of the ranked list $\mathscr{T}$. Thus, in this phase, we obtain the preliminary ranked list of influential nodes $\mathscr{T}$ of size α by only relying on the temporal sequence of retweets.

Step 3: Final ranked list of influencers $\mathscr{S}$.

In the second phase, we exploit the available follower information of users in $\in \mathscr{T}$ in order to obtain a refined ranked list of influential nodes $\mathscr{S}$ of size $\beta \leq \alpha$. The idea of this refinement is to filter out users from $\mathscr{T}$ who have high follower overlap and only keep one of them, so that users in $\mathscr{S}$ have maximal neighbours' diversity, as the initial diversity of a cascade across several parts of the network favours its spreading [Weng et al. 2014]. We implement a variation of the set cover problem [Caprara et al.

2000] that more generally aims at covering all the elements of a universal set using a minimum number of available subsets. In our context, the universal set corresponds to all the users and each available subset corresponds to a user and its followers. The problem consists in checking whether the users, picked in the order of their position in the ranked list $\mathscr{T}$, allow the content to be exposed to a new audience. Typically, these methods require an initial random list that gets refined. In our situation, the choice of the initial list will be $\mathscr{T}$ rather than a random one. Following the ranked list $\mathscr{T}$ at each step, we incrementally populate the list $\mathscr{S}$ by checking whether the corresponding influencer u in $\mathscr{T}$ exposes the content to at least one of its followers. This ensures that influential nodes in the refined list $\mathscr{S}$ expose the content to a diverse population in the network.

In a nutshell, the SmartInf algorithm may be summarized in three steps as follows:

1. Detect the first Peak in Type I cascades;
2. Rank the users according to their number of retweet around the first peaks;
3. Filter the ranked list in order to ensure enough neighbours'diversity.

7.2.3 Computational complexity

Both the detection of the peaks and the identification of potential influential users belonging to the set $\sqcup$ are of complexity $O(|\mathbf{T}|)$ where $\mathbf{T}$ is the set of Type I cascades. The algorithm *SmartInf-Temp* computes the frequency of appearance of the users in $\sqcup$. For each cascade $C \in \mathbf{T}$, it updates the frequency of appearance for the v^C users retweeting between the first peak and the subsequent fall. Thus, the time required for updating this frequency across all cascades in $\mathbf{T}$ is $v^C|\mathbf{T}| \sim O(|\mathbf{T}|)$. Then, sorting the users in $\sqcup$ to obtain $\mathscr{T}$ has the complexity $O(\alpha \log \alpha) \sim O(N \log N)$ since $|\mathscr{T}| = \alpha$ and $\alpha \leq N$. Thus, the total time complexity of Algorithm *SmartInf-Temp* is $O(|\mathbf{T}|) + O(N \log N)$. The second phase of *SmartInf* checks whether each user in the ranked list $\mathscr{T}$ exposes any new follower with complexity $O(\alpha) \sim O(N)$. So the overall time complexity of *SmartInf* is $O(|\mathbf{T}|) + O(N \log N)$. Essentially, *SmartInf* does not require to directly measure the influence of each node in the whole network, which can be costly, in order to rank influential nodes. Moreover, a maximum of α lists of followers of well-selected users (from $\sqcup$) are required, which also allows to be economical in crawling the network's structure, instead of blindly crawling the global network.

7.2.4 Importance of the refining step

The refinement of the ranked influencer list $\mathscr{T}$ to obtain $\mathscr{S}$ enables the tweet to spread to a more diverse population. In order to demonstrate that this refinement step is useful, we compute the average pairwise overlap among the top-k influencers obtained from *SmartInf-Temp* ($\mathscr{T}$) and *SmartInf* ($\mathscr{S}$) in Figure7.1 for *Algeria* (left) and *Egypt* (right) datasets. In both datasets, we observe that the mutual overlap is much higher in case of *SmartInf-Temp* compared to *SmartInf*, which indicates that the refinement step actually enables to obtain an influencer list that reaches a more diverse population.

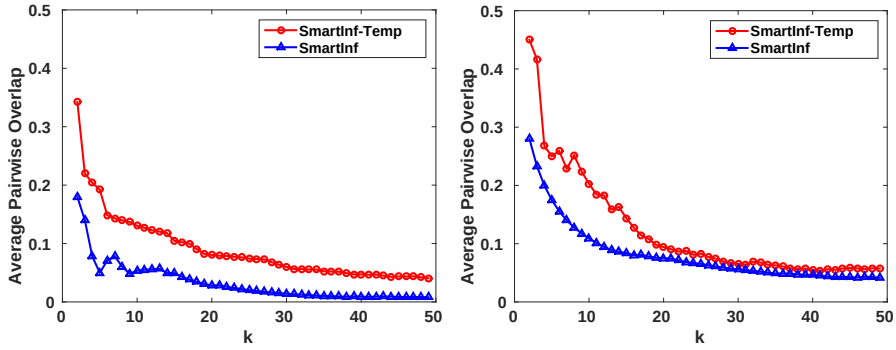

Figure 7.1 – The refining step of *SmartInf* actually allows to decrease the pairwise follower overlap of the top-k users of the list $\mathscr{S}$ obtained from *SmartInf* compared to the list $\mathscr{T}$ obtained from *SmartInf-Temp* applied to the *Algeria* [left] and *Egypt* [right] datasets.

7.2.5 Standard centrality measures

The classical centrality measures provide a first indication about the quality of a set of influential nodes. In Table 7.1 we provide the average centrality scores of the top-10 users recommended by *SmartInf* against the top-10 of each studied centrality measures and against 10 randomly selected users (averaged over 500 samples). We consider the three following standard centrality measures: Pagerank centrality (**PR**) [Ding et al. 2009], Eigenvector centrality (**EV**) [Ruhnau 2000] and Betweenness centrality (**BW**) [Riquelme and González-Cantergiani 2016]. These metrics essentially capture the structural centrality of the influential nodes.

Algorithm	**PR** (10^{-5})	**EV** (10^{-5})	**BW** (10^{-5})
SmartInf (*Algeria*)	53.9	600	100
Random (*Algeria*)	0.583	10	0.0001
Top-10 (*Algeria*)	100	6000	640
SmartInf (*Egypt*)	1.96	78.6	9.83
Random (*Egypt*)	0.0002	0.000002	0.00003
Top-10 (*Egypt*)	50	900	120
SmartInf (*Nepal*)	8.9	40	400
Random (*Nepal*)	0.031	0.0188	20
Top-10 (*Nepal*)	40	620	1200

Table 7.1 – Comparison of mean score of the classical centrality metrics for top-10 influential nodes recommended by *SmartInf*, randomly selected top-10 nodes and the top-10 nodes of each metrics. *SmartInf* provides users with decent centrality scores.

We observe that the top-10 influential nodes identified by *SmartInf* are not the nodes with the highest centrality scores but they exhibit higher centrality scores compared to the randomly selected nodes consistently across *Algeria*, *Egypt* and *Nepal* datasets. Close investigation reveals that the top-10 structurally central nodes according to a given metric either remain inactive in posting retweets or mostly participate in non-popular cascades. Thus, those central nodes may not play a key role in the migration of the cascade to a different diffusion locality. Consequently, although they possess a structurally favourable position in the follower network, they are not recommended by *SmartInf*. This implies that although *SmartInf* does not explicitly filter the nodes with the highest centrality scores, it naturally selects active enough nodes which exhibit decently high centrality scores.

We considered these basic centrality measures as first performance indicators. A more detailed analysis could have consisted in computing the $k-$score (purely structural), the TempoRank score [Rocha and Masuda 2014] (that allows one to combine structural position and overall users'activity) or the burstiness of the reaction time distribution of the users (which may indicate a tendency to trigger cascades). However the Twitter-specific metrics as well as the epidemic simulations studied in the next section already provide complementary performance measures.

7.3 Evaluation of SmartInf performance on empirical data

We will now compare the performance of the *SmartInf* and *SmartInf-Temp* algorithms against various baseline algorithms.

7.3.1 Baseline algorithms

We consider the following competing algorithms for identifying the set of influential users in a network:

1. **K-truss decomposition:** This [Malliaros et al. 2016] is a triangle-based extension of the k-core decomposition of graphs that extracts a denser subgraph compared to k-core [Carmi et al. 2007]; it is structurally closer to a clique achieving faster and wider epidemic spreading.

2. **MCDWE score:** This method [Sheikhahmadi and Nematbakhsh 2017] combines the core number of a node, its degree and its neighbours' diversity based on k-shell decomposition to rank the users.

3. **MCDWE-Activity:** This is a *hybrid method* that combines the activity rate and the structural information based on the MCDWE scores [Sheikhahmadi and Nematbakhsh 2017] of the users. In this method, we simply compute the product of frequency of retweets and MCDWE scores of all users; we then sort users in decreasing order of this product to get the ranking.

We tested various other baseline algorithms including *UIRank* [Jianqiang et al. 2017], *Collective influence* [Morone et al. 2016] and *Meta-centrality* [Madotto and Liu 2016], but we only kept the aforementioned ones as they were consistently providing the best results in terms of used metrics and numerical simulations.

7.3.2 Twitter-specific metrics

We first consider some Twitter-specific metrics in order to compare the lists of influencers suggested by *SmartInf* and the baselines.

7.3.2.1 Definition of the Twitter-specific metrics

(a) Gain in retweet count $\mathscr{R}^u$

This metric measures the impact of retweets of a given user (say u) on the final size of the cascade [Ding et al. 2013; Riquelme and González-Cantergiani 2016; Al-Garadi et al. 2018]. Consider a cascade C of size n_C where the k_C^{th} retweet has been posted by the user u_k^C. Without any prior knowledge of the structure of the underlying follower network, each of the $n_C - k_C$ new retweets occurring after the k_C^{th} retweet is equally contributed by the first k_C users who have already retweeted in C.

Thus, the relative gain in size of the cascade C due to the retweet posted by user u_k^C is defined as $\frac{n_C - k_C}{n_C k_C}$. Finally, the *Gain in retweet count* $\mathscr{R}^u$ for the user u is computed as the average gain over all the cascades (N^u) in which u participates, which leads to

$$\mathscr{R}^u = \frac{1}{|N^u|} \sum_{C \in N^u} \frac{n_C - k_C}{n_C k_C}. \tag{7.3.1}$$

Thus, this metric, which is agnostic to the underlying follower network structure, favours users retweeting in the early stage of large cascades compared to the users involved in small cascades, or participating in the later stage of the large cascades. These favoured users act as precursors of the popularity of the tweet; they should therefore be preferentially targeted for viral spreading.

(b) Gain in exposed user count $\mathscr{E}^u$

This metric measures the volume of newly exposed population a_C, who retweeted the content in the cascade C only due to retweet posted by a specific user u [Ding et al. 2013; Riquelme and González-Cantergiani 2016]. The exposure to the content may be caused directly or indirectly via the follower links of u.

Similarly to the Gain in retweet count $\mathscr{R}^u$, the *Gain in exposed user count* $\mathscr{E}^u$ for user u over the M^u cascades in which u participated can be expressed as

$$\mathscr{E}^u = \frac{1}{|N^u|} \sum_{C \in N^u} \frac{a_C}{k_C n_C}. \tag{7.3.2}$$

This metric scores high for the users who tend to be responsible for the first exposure of the tweet to many of the retweeting users in a cascade. Therefore, this metric reflects their capacity to expose the content to a new and relevant audience.

(c) Active gain in retweet count and exposed user count $\mathscr{I}^u$ and $\mathscr{H}^u$

These metrics take the activity (retweet) rate $\mathscr{P}^u$ of user u into consideration and combines it along with the aforesaid two metrics $\mathscr{R}^u$ and $\mathscr{E}^u$ respectively. Precisely, we define

$$\mathscr{I}^u = \mathscr{P}^u \mathscr{R}^u \tag{7.3.3}$$

$$\mathscr{H}^u = \mathscr{P}^u \mathscr{E}^u, \tag{7.3.4}$$

where the retweet activity $\mathscr{P}^u$ is estimated empirically from the cascades in which u participated by computing how many times the user u retweeted a tweet it got exposed to. In principle, these metrics score high for the users who are more likely to retweet the tweet content and have high gains $\mathscr{R}^u$ and $\mathscr{E}^u$. The combination of these two metrics allows to discard the users with high gains but who tend to participate less to the cascades, as well as those who tend to retweet a lot but typically fail to propagate the tweet efficiently.

7.3.2.2 Evaluation based on the Twitter-specific metrics

We obtain the top-k recommended influential nodes $\mathscr{S}^k$ from *SmartInf* as well as from the baseline algorithms. We report the average score of the Twitter-specific metrics of the corresponding sets for $k = 10, 20$ and 50 in Tables 7.2, 7.3 and 7.4 for *Algeria*, *Egypt* and *Nepal* datasets respectively.

We observe that both the *Gain in retweet count* $\mathscr{R}^u$ and *Gain in exposed user count* $\mathscr{E}^u$ attain highest average scores for top-k influential nodes $\mathscr{S}^k$ identified by *SmartInf* (and the variation *SmartInf-Temp*) across each datasets. This indicates that influential users detected by *SmartInf* and *SmartInf-Temp* mostly retweet in the early part of large cascades, as well as are responsible for exposing the tweet to a large fraction of retweeting users compared to the influential nodes identified by the baselines.

Since the baseline algorithms (*MCDWE* and *K-truss*) mostly rely on the network structure, the average scores of $\mathscr{R}^u$ and $\mathscr{E}^u$ are low because influential nodes identified by such methods either retweet in small cascades or towards the later part of large cascades. Moreover, some of these influential nodes do not even participate in retweeting activity. Hence, the values of $\mathscr{I}^u$ and $\mathscr{H}^u$ are very poor for the baselines.

Notably, *MCDWE-Activity* performs better than *MCDWE* and *K-truss* across all the metrics since it combines the retweet activity of users with structural information. However, it performs worse than *SmartInf* and *SmartInf-Temp* since corresponding influential nodes with high retweet rate in case of *MCDWE-Activity* do not retweet early in large cascades and expose the tweet to a limited fraction of retweeting users.

These results highlight the fact that the standard baselines fail in taking into account the behavourial heterogeneity among the users, which is an important feature to take into account when considering influential users.

Number of seeds	k=10				k=20				k=50			
Algorithm	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$
SmartInf	**0.480**	**0.054**	**0.350**	0.036	**0.470**	0.036	**0.400**	0.030	**0.450**	0.017	**0.360**	0.010
SmartInf-Temp	0.450	0.041	0.330	**0.040**	0.420	**0.066**	0.320	**0.050**	0.420	**0.036**	0.280	**0.040**
MCDWE	0.080	0.003	0.090	0.002	0.090	0.004	0.070	0.003	0.060	0.002	0.050	0.002
K-truss	0.226	0.030	0.100	0.006	0.230	0.010	0.120	0.005	0.190	0.006	0.110	0.003
MCDWE-Activity	0.360	0.037	0.280	0.028	0.310	0.028	0.150	0.014	0.220	0.013	0.100	0.006

Table 7.2 – Mean score of the influence metrics for sets of top-k influential nodes taking $k = 10, 20, 50$ on the *Algeria* dataset.

Number of seeds	k=10				k=20				k=50			
Algorithm	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$
SmartInf	**0.405**	**0.210**	**0.300**	**0.092**	**0.340**	0.040	**0.310**	0.010	0.350	0.020	**0.300**	0.010
SmartInf-Temp	0.318	0.058	0.261	0.006	0.310	**0.200**	0.260	**0.080**	**0.390**	**0.150**	0.270	**0.060**
MCDWE	0.064	0.002	0.006	0.002	0.040	0.001	0.010	0.001	0.020	0.009	0.010	0.004
K-truss	0.122	0.005	0.060	0.004	0.110	0.003	0.050	0.002	0.120	0.004	0.040	0.001
MCDWE-Activity	0.167	0.009	0.022	0.012	0.160	0.007	0.080	0.004	0.150	0.010	0.070	0.005

Table 7.3 – Mean score of the influence metrics for sets of top-k influential nodes taking $k = 10, 20, 50$ on the *Egypt* dataset.

Number of seeds	k=10				k=20				k=50			
Algorithm	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$	$\mathscr{R}^u$	$\mathscr{I}^u$	$\mathscr{E}^u$	$\mathscr{H}^u$
SmartInf	**0.340**	**0.030**	**0.200**	**0.020**	**0.270**	**0.030**	**0.180**	**0.020**	**0.250**	0.020	**0.160**	**0.020**
SmartInf-Temp	0.220	0.020	0.140	0.010	0.230	**0.030**	0.140	**0.020**	0.230	**0.030**	0.140	**0.020**
MCDWE	0.170	0.0009	0.150	0.0004	0.150	0.006	0.120	0.004	0.160	0.006	0.140	0.004
K-truss	0.210	0.006	0.130	0.004	0.180	0.008	0.110	0.005	0.190	0.007	0.120	0.004
MCDWE-Activity	0.200	0.015	0.140	0.013	0.210	0.016	0.140	0.010	0.200	0.018	0.140	0.014

Table 7.4 – Mean score of the influence metrics for sets of top-k influential nodes taking $k = 10, 20, 50$ on the *Nepal* dataset.

7.3.3 Epidemic simulation

The Twitter-specific metrics allowed us to indirectly evaluate the quality of the different sets of influential users provided by the different algorithms. However, the goal of these algorithms is to provide a set of influential users maximizing the spread of a tweet. A more direct evaluation of the quality of these algorithms consists in evaluating the actual size of multi-seeded cascades when the seed set corresponds to the one suggested by the different algorithms. Such cascades do not exist in our datasets, and a real-world experiment is practically impossible to conduct. Thus, we turn to numerical simulations on the empirical networks in order to evaluate the final size of such multi-seeded cascade.

We implement the standard susceptible-infected (SI) model of epidemic simulation introduced in section 1.6.1 to evaluate the quality of a set of influential nodes in the network obtained in the *Algeria* and *Egypt* datasets. The transition of the seeds from susceptible (S) to infected state (I) indicates that the user retweets the initial post. Once a susceptible node in the network becomes infected, its followers get exposed to the post and get a single chance to change their state to the infected one, depending on their probability to retweet β_u. Therefore, we discard the re-exposure phenomenon. The infection probability β_u varies across the users $u \in U$ in the network and is com-

puted empirically as the fraction of cascades in which u retweeted over all the cascades where u got exposed to the tweet.

The initial seed is an artificial external node of the network whose neighbours are the recommended set of users which we refer to as seed nodes. This corresponds to the situation where an eternal source wants to propagate a tweet by selecting a given number of users to target.

The average volume of the final infected population over 1000 realizations for varying number (top-k) of seed nodes $\mathscr{S}^k$ is provided in Figure 7.2 for *Algeria* and *Egypt* datasets.

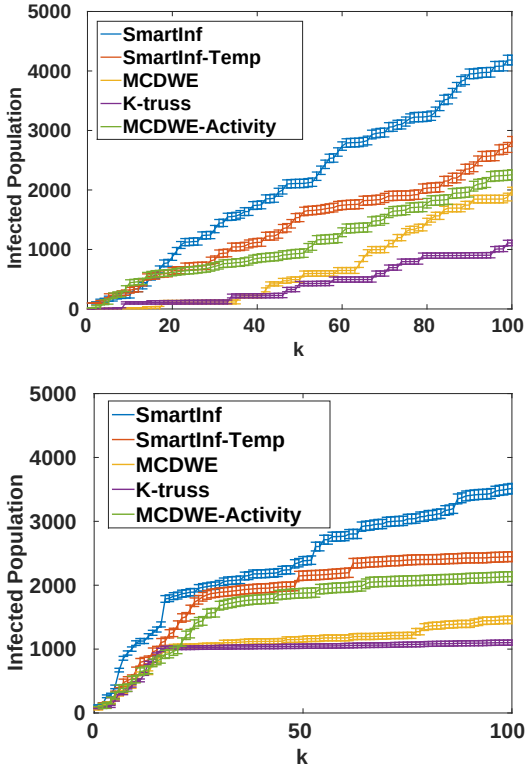

Figure 7.2 – Average final cascade size for varying number of seeds corresponing to the top-k influential nodes recommended by the different algorithms (errorbars indicate 2 standard deviations) on networks corresponding to *Algeria*[Top] and *Egypt* [Bottom] datasets.

We observe that the seed nodes corresponding to *SmartInf* achieve the largest infected population compared to the baselines. This result demonstrates the spreading capacity of the influential nodes identified by the proposed *SmartInf* algorithm. As we increase the number k of seed nodes, we observe that the final volume of the infected population increases quite steadily for *SmartInf*. On the other hand, for baselines,

the volume of infected population increases slowly for smaller values of k and then saturates for larger k indicating that there is an increasing overlap in the exposed population when the number of seeds increases. We can also observe the improvement in performance of *SmartInf* over its variation *SmartInf-Temp*; this again depicts the significance of the refinement phase that minimizes the overlap in population infected by different seeds, thus reaching out to a diverse audience.

7.3.4 Quality of influential nodes obtained by Smartinf

We investigate several key features of the detected top-k influential nodes $\mathscr{S}^k$ using *SmartInf* that we compare to the baselines as shown in Figure 7.3 (for $k = 50$) on the empirical datasets.

We observe that *SmartInf* identifies influential nodes with a higher retweet rate than the ones from *MCDWE* and *K-truss* but lower than the ones from *MCDWE-Activity*, which is expected as this latter method explicitly takes retweet frequency into consideration. In a similar vein, we observe that the influential nodes of *SmartInf* have a decent follower count that is higher than *MCDWE-Activity* (though lower than *MCDWE* and *K-truss* which primarily relies on identifying high degree nodes belonging to the core of the network).

Nevertheless, *SmartInf* influencers expose the post to a larger mew audience than the ones from the baselines. Finally, we observe that the mean reaction time is also the lowest for the influencers of *SmartInf*. Notably, the influential nodes $\mathscr{S}$ identified by *SmartInf* respect a balance between a good activity and a strategical position in the network, whereas *SmartInf* does not explicitly consider the topological structure or the retweeting behavior of the users.

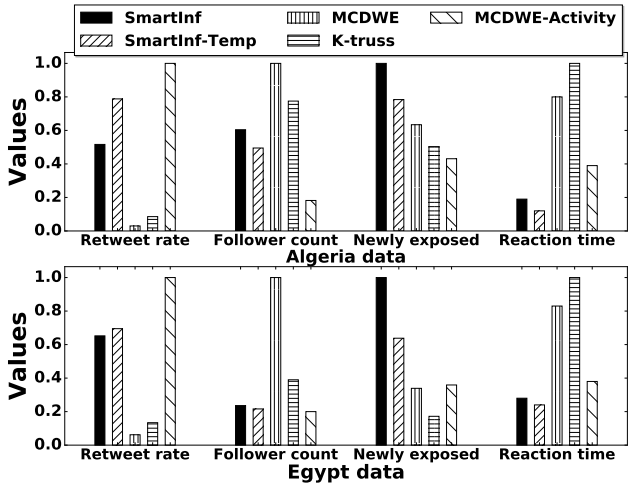

Figure 7.3 – Evaluation of *SmartInf* top-50 users compared to the baselines based on several key features.

7.3.5 Robustness of SmartInf

Finally, we investigate the robustness of the influential nodes $\mathscr{S}$ obtained from *SmartInf* against the volume of available Type I cascades. Indeed, the principle of *SmartInf* relies on the Type I cascades and we want to determine whether our method provides a similar ranking of influential users when removing a given fraction of Type I cascades from the empirical data. Since the method detects influential users with the highest activity around the first peaks of Type I cascades, we investigate whether such users continue to have a high activity around the peaks in the case of missing data.

In Figure 7.4, we compute the Spearman rank correlation [Gauthier 2001] between the top-k influential nodes $\mathscr{S}^k$ (for $k = 50$) identified by *SmartInf* using the entire dataset and the corresponding set $\mathscr{S}_r^k$ of influential nodes when a fraction r of Type I cascades have been removed from the *Algeria* dataset. We compare this Spearman correlation to the one between $\mathscr{S}^k$ and the different sets provided by the baselines. We observe that $\mathscr{S}^k$ and $\mathscr{S}_r^k$ are more correlated to each other than $\mathscr{S}^k$ and the baselines when removing up to 40% of the data. We also observe that the rank correlation remains quite large (0.34) even on removal of 50% of Type I cascades , depicting the robustness of *SmartInf* even in the face of missing data.

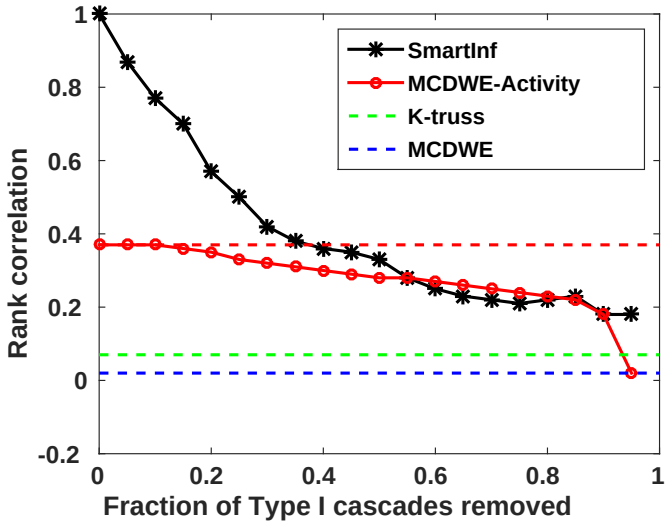

Figure 7.4 – Rank correlation of the list provided by *SmartInf* with the full data against the rank provided when removing a fraction of Type I cascades in *Algeria* dataset.

7.4 Evaluation of SmartInf performance on synthetic data

Finally, we generate synthetic cascades on an artificial network in order to highlight some additional key factors behind the performance of *SmartInf*. We will generate cascades that have some properties similar to the ones in the empirical datasets, in a network for which we will exactly know the structure.

7.4.1 Synthetic setup

7.4.1.1 Synthetic network

We construct a synthetic network based on the stochastic block model [Holland et al. 1983] with two blocks. Such a network exhibits a natural partition of nodes through its blocks which inherently correspond to two different diffusion localities. We construct a directed network composed of two Erdős-Rényi (E-R) [Abbe 2017] sub-networks E_1 and E_2 with parameters (n_1, n_2, p_1, p_2, q) where n_i and p_i denote the size and density of the block E_i, whereas q is the inter-block density. We call *bridges* the m links connecting the two blocks and *bridge nodes* their extremities. We choose Erdős-Rényi blocks because the degree distribution is Poissonian and therefore no hub emerges. Therefore, in a purely structural perspective, the influential nodes of the network lie among the bridge nodes. We have $m \approx n_1 n_2 q$ and set $n_1 n_2 q \ll \min(n_1^2 p_1, n_2^2 p_2)$ [Abbe 2017], ensuring the natural partition of the network into two localities between which spreading of tweets may occur. In our simulation, we fix the parameter values as $n_1 = 1000$, $n_2 = 1000$, $p_1 = 0.2$, $p_2 = 0.3$ and $q = 5 * 10^{-5}$ such that $m \approx n_1 n_2 q = 50$.

7.4.1.2 Generation of the synthetic cascades

Following a similar method as in section 7.3.3, we generate the synthetic cascades using a standard SI model, where initially all but one randomly selected seed u_0 are in the susceptible state. The followers of u_0 get exposed to the post and retweet it with probability β_u.

The time t_i of the retweet of the user u_i is computed from its reaction time μ_i which corresponds to the elapsed time between the moment u_i got exposed to the tweet for the first time and the moment it retweeted it. Consistently with previous studies [Artime et al. 2017], our empirical analysis reveals that the reaction time depends on whether u_i and u_j, the user who exposed the tweet to u_i, belong to the same block or to different blocks. In Figure 7.5, the empirical complementary cumulative distribution functions (ccdf) on the *Algeria* dataset show that the distributions f_{intra} of intra-block reaction times and f_{inter} of the inter-block reaction time significantly differ as f_{inter} is broader and has a heavier tail than f_{intra}. This distinction favours a faster spread of the tweet inside a diffusion locality. Thus, while forming the synthetic cascades, we draw the reaction times μ_i of a retweeting user u_i from the suitable reaction time distributions and we subsequently compute the retweet time instance t_i^D. In total, we generate 10000 cascades of mean size 327, the largest being of size 476, which are

similar values as in the studied datasets. Moreover, the reaction time distributions are heavy-tailed, which is a condition that facilitates the existence of Type I cascades, as we analytically discussed in section 6.4.1.

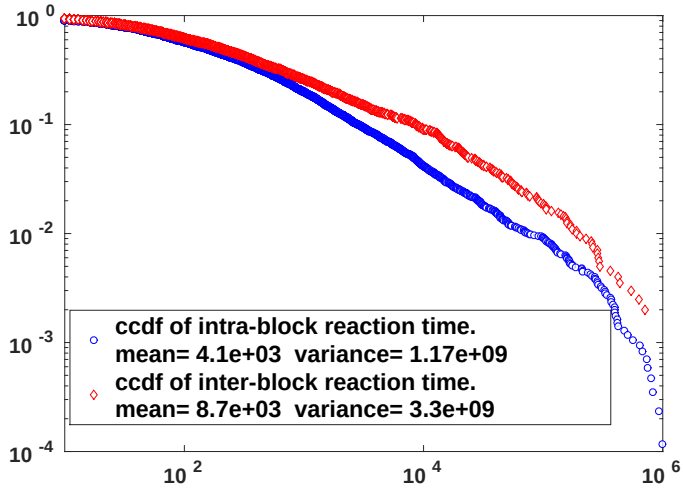

Figure 7.5 – CCDF distribution for intra-block and inter-block reaction times from the *Algeria* dataset. The latter is significantly broader.

7.4.1.3 Properties of the synthetic cascades

The model has a lot of parameters. Among all possible values, we need to select a set of parameters that provides a sample of cascades with properties similar to the one of our dataset and a network with the appropriate structure. In particular, we require:

1. the existence of two distinct blocks;
2. the existence of Type I cascades ;
3. the saturation and migration phenomenon behind the emergence of Type I cascades.

We first vary the number m of bridges to allow the emergence of enough Type I cascades. A too small m restricts the tweet from spreading to the second block whereas a too large m allows the tweet to quickly spread to the second block before the saturation of the first block. In both cases, the resulting cascades are of Type II. For the obtained Type I cascades, we check whether the temporal peaks in Type I cascades co-occur with the migration of the cascade to a new diffusion locality, as displayed in Figure 7.6. Precisely, we consider that a migration occurs if at least 95% of the retweeting users before the peak belong to the first block and at least 80% of the retweeting users after the peak belong to the second block. In Figure 7.7, we highlight the fact that such migration only occurs for the generated Type I cascades, similarly to the findings

in the empirical datasets of section 6.5.1. We finally keep 50 bridges for two reasons. On the one hand, it allows to generate a decent proportion (53%) of Type I cascades among all the cascades. On the other hand, a large proportion (80%) of such Type I cascades results from of a saturation/migration phenomenon. Moreover these proportions are in the range of the ones previously empirically observed in the datasets (Table 6.2 in section 6.3.1 and section 6.5.1).

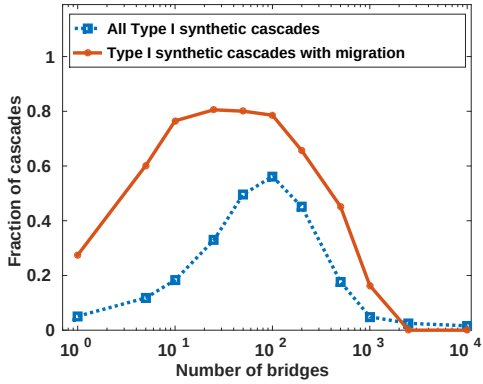

Figure 7.6 – Fraction of Type I synthetic cascades obtained for different number m of bridges. The selection of $m = 50$ allows to have a significant proportion of cascades of Type I among all the cascades and a significant amount of cascades of Type I for which the first peak corresponds to the migration of the synthetic cascades.

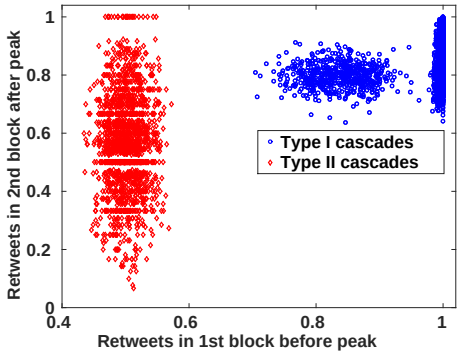

Figure 7.7 – Proportion of retweets across different blocks before and after the first peak for $m = 50$. Only the obtained Type I cascades show migration between the two blocks after the first peak, as observed in the real data.

7.4.2 Experimental observations

We rely on the synthetic dataset to investigate the factors behind the superior performance of the proposed *SmartInf* algorithm.

We investigate in Figure 7.8 how the different methods perform in detecting the bridge nodes. We observe that a majority of bridge nodes emerge in the set of top-k influential nodes $\mathscr{S}^k$ selected by *SmartInf* without any explicit utilization of the structural information. The bridge nodes play an important role in the spreading of a cascade across multiple localities. On the contrary, the percentage of bridge nodes is low among the top-k influential nodes detected by the baselines; the reason is that influential nodes identified by the baselines (*MCDWE* and *K-truss*) belong to the core of the network and therefore occupy central positions within a single diffusion locality of the network.

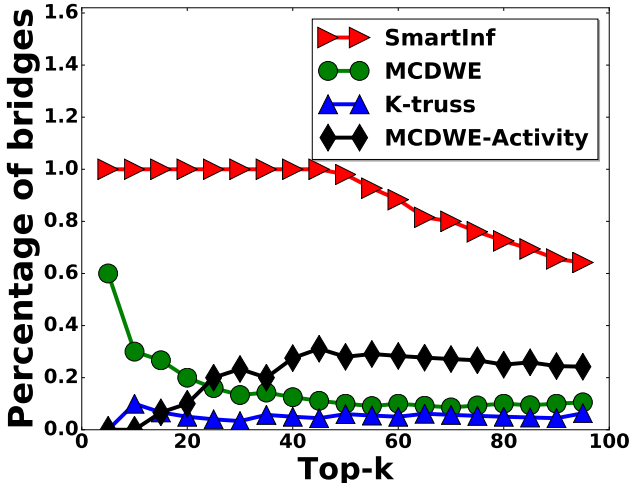

Figure 7.8 – Proportion of bridge nodes in top-k of the different algorithms.

Additionally, the reaction time μ also plays an important role in the speed of the spreading of the tweet to a new locality. Indeed, a user retweeting faster the information coming from another part of the network helps the tweet to propagate faster. We will show that such user may be detected by *SmartInf*.

We focus on the bridging nodes and associate to each bridge node u a reaction factor s_u. The reaction time μ_u of a user u follows the distribution $s_u f_{\text{inter}}$ to retweet an information coming from a different block, and $s'_u f_{\text{intra}}$ to retweet an information coming from the same block, where s'_u is defined so that the mean reaction time of a user is constant for each bridge node u. The smaller reaction factor s_u of u indicates the faster migration of the tweet to the second block. Therefore, it is a sign of qual-

ity for a method to rank better users with lower reaction rate s_u. However, as both the mean retweet time and the retweet rate of a bridge node remains constant, all the baselines, including *MCDWE-Activity*, will exactly provide the same ranking than in the previous case where reaction factor was set to 1 for each bridge node. Therefore, only *SmartInf* may be sensitive to the reaction factor s_u.

Generating cascades with varying reaction factors s_u for different bridge nodes shows that the Spearman correlation coefficient between the ranking of the influential nodes provided by *SmartInf* and the reaction factor s_u is large (0.76), whereas it is lower for the baseline algorithms (maximum 0.28). Thus, *SmartInf* is able to detect this key feature enabling faster spread, where the baselines fundamentally fail to do so.

7.5 Discussion

This work addresses the important problem of identifying a set of influential users in a social network by leveraging on the temporal sequences of retweets of Twitter cascades. To this end, we present *SmartInf*, an algorithm leveraging on *anchor nodes* whose retweets can expose the cascade to a large new population. The ranked list of influential nodes obtained from temporal retweet sequences is then refined to ensure that a diverse population can be reached through influential nodes identified by *SmartInf*. Importantly, *SmartInf* requires alternative and typically cheaper inputs than the standard baselines. It requires a history of similar cascades that previously spread over the network and the availability of the follower list of specific users, whereas the baselines typically require the full structure of the underlying network. Even though *SmartInf* only relies on Type I cascades, our experiments on various datasets have shown that it is very likely that a significant proportion of the cascades from a given Twitter dataset are of Type I. Therefore, the requirement of the existence of Type I cascades for *SmartInf* to be applied to a dataset does not seem to be a heavy limitation to its use.

We have demonstrated through experiments on multiple empirical datasets from Twitter that *SmartInf* achieves a significant performance boost over the recent state-of-the-art baselines, both in terms of proposed influence metrics and in terms of the volume of infected population using epidemic simulation. We have also shown that *SmartInf* is robust to missing data and scales linearly with the size of the network. Finally, we have investigated the factors behind the superior performance of *SmartInf* on a simple synthetic network through simulation of synthetic cascades on this network. Our experiments have revealed that *SmartInf* can detect such influential nodes which play bridging roles in the network, connecting multiple localities as well as enable a faster spread of information. All these results point to the fact that our proposed *SmartInf* algorithm provides a set of influential users that can quickly spread the message to a large and diverse population in the network.

Our present work has multiple future research directions. First, one may be curious to investigate whether *SmartInf* recommended influential nodes vary significantly across the different topics (such as politics, sports, entertainment etc.) in Twitter posts, which would require new comprehensive Twitter datasets. Second, there are several threshold to set for the algorithm, specifically for the length of the window around the peak and during the refinement step. It would be more elegant to learn them in an automated manner, for instance based on features extracted from inter-retweet intervals. Third, further experiments on synthetic cascades with richer structural properties may provide deeper insights on the efficiency of *SmartInf*. Fourth, we have completely overlooked the role of Type II cascades for influential users detection because *SmartInf* only relies on Type I cascades. It would be an interesting research direction to exploit the Type II cascades in order to refine the list of influencers obtained from *SmartInf*. For instance, one might discard users who tend to retweet during the end of Type II cascades, as they failed to propagate the tweet to a new population when the current locality was saturated. On the contrary, mainly retweeting in the early phases of Type II cascades might indicate that the user helps the tweet to propagate before it saturated the current diffusion locality, which is in particular interesting for short-life content. Finally, we have suggested to use *SmartInf-Temp* when the underlying structure is not available. A further research direction could consist in considering the alternative solution of estimating the network structure through standard inference methods, such as [Gomez Rodriguez et al. 2016; Rodriguez et al. 2013], and apply *SmartInf* on the inferred network.

Discussion

This thesis aimed at developing theoretical models and practical methods for temporal networks. The objective was two-fold.

First, we studied simple stochastic processes such as Random Walks on temporal networks. Despite their simplicity, the models have demonstrated the impact of a bursty activity on the resulting trajectory. Our findings emphasize the importance of the temporal patterns of the entities involved in the process, in supplement to the classical topological approach. This temporality may be of two different kinds. The first one lies in the temporality of the diffusing entity, which may have its own internal timescales. For instance, the entity may born, die, incubate or need some time to rest. The second reason behind the temporal patterns is the dynamics of the network on which the entity diffuses. Agents may not be always active or some connections may be available only for some given duration. We considered several models of Random Walks and showed that temporal-induced biases in the resulting trajectory may emerge, in particular due to the bursty behaviour and the presence of short cycles on the underlying network. The main contribution of this first part is the exhibition of such biases in various types of Random Walks on temporal networks. Importantly, this implies that purely temporal properties impact the structural exploration of the network.

Future research directions include exploring the impact of these biases on the trajectory of a Random Walker on methods providing metrics for large networks, for instance in many community detection methods. Since they tend to capture probability flow, one question that arises is whether those short cycles are over-represented compared to Markovian dynamics, and to what extent. We focused on the biases due to the smaller cycles, namely the backtracking for undirected network and the cycles of length two for directed network, because they intuitively induce larger bias on the trajectory of a Random Walker compared to longer cycles. However, we did not quantify how large this effect is compared to other short-cycles, and in particular to triangles. Such an analysis may be of interest for future work.

Second, we studied spreading models of information or disease in order to adapt the spreading strategy under bursty behaviour. In a similar fashion as in the study of Random Walks, we have shown that temporal patterns associated to a cascade that diffused on Twitter, in particular its inter-retweet time intervals, may provide information about how it spread on the network. Combined together, our findings show that one may significantly increase the final share of a message by focusing on two key aspects: improving its quality rather than spamming it on the one hand and targeting a well-selected set of users on the other hand. These results hold assuming that the transmissibility does not change over time, which is however not always guaranteed and would deserve further exploration.

We developed the algorithm *SmartInf* that provides such a set of users whose specific targeting allows to increase the popularity of the message. One of the main advantage of *SmartInf* is that it does not require the topology of the underlying network which is typically costly to get. Alternatively, it requires a set of cascades that historically spread on the network. Thus, *SmartInf* is particularly well adapted to information diffusion on social networks such as Twitter or Facebook where many cascades occurred but not to disease spreading on physical networks. The list of targets may be improved if a small number of ego-networks of some specific users are available. Moreover, *SmartInf* does not need to compute a score for all the users, which makes it faster than the baseline methods. *SmartInf* also allows one to obtain a list of targets preserving the privacy of the specific interactions between the users.

SmartInf may be improved in several aspects, among which the fine-tuning of its several thresholds and the non-exploitation of the Type II cascades. However, the main result about *SmartInf* is that combining temporal and structural information provides a better result than solely using the latter, rather than the exact performance of the algorithm itself. Moreover, *SmartInf* requires less structural information than the standard baselines. As a consequence, the required amount of interaction data between the users is reduced as it only focuses on some specific influential users. Also, it is worth mentioning that our early attempts using several machine learning techniques (support vector machines, neural networks) provided worse results than the ones obtained using the baseline algorithms. We did not investigate further into these directions as our more classical approach allows to have an understanding of the reasons why it performs better than the baseline algorithms.

One aspect we neglected in our study of retweets is the lifetime of the spanned cascade. Indeed, as time passes by, the content of a message gets diffused through other media and through other competing similar messages, reducing its potential of reshare: from being a scoop, the message evolves to becoming a standard information, then history. Consequently, it may not be enough to be certain that the initial tweet will be shared to a large audience, it needs to reach it fast. Simulating the message spreading through an SIR model seems to be the most natural extension to incorporate such time limitations.

A further research direction in exploiting the time series of cascades concerns the reconstruction of networks based on temporal patterns. Such inferred networks based on the dynamics differ from the initial physical one as they aim at reproducing influence relations. Indeed, because of the heterogeneity of the behaviours, the function of the users varies, and the underlying static network is only a proxy of the network of influence that determines the spreading dynamics. The problem is that there exists no ground truth for this influence dynamics, and performance may only be evaluated through metrics of diffusion. An interesting approach would be to compare the ranking obtained for the set of users suggested by *SmartInf* with topological metrics obtained from several reconstruction network methods. If we detect users that are ranked high with independent methods, this would provide more confidence in their importance in terms of message diffusion.

Finally, a societal issue in online social networks that arose in the last years concerns the massive spread of fake news, as it came to light during the 2016 US elections. Preliminary efforts have been made in the last years on the understanding and detection of fake or unverified facts. It is of societal importance that the scientific community keeps on improving bots and fake news detection methods, and the analysis of their peculiar associated temporal behaviour is a promising direction. This last example emphasizes the importance of the information and communications technologies (ICT) in the society from a world-wide perspective. The potential of development in the big data era is huge, from health care and education to space exploration and biological interactions, but it requires strong safeguards as the boundary between public and private life increasingly blurs.

Bibliography

E. Abbe. Community detection and stochastic block models: recent developments. *JMLR*, 18(1):6446–6531, 2017.

M. A. Al-garadi, K. D. Varathan, and S. D. Ravana. Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. *Physica A*, 468:278–288, 2017.

M. A. Al-Garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, G. Mujtaba, M. U. S. Khan, and S. U. Khan. Analysis of Online Social Network Connections for Identification of Influential Users: Survey and Open Research Issues. *ACS*, 51(1):16, 2018.

D. Aldous and J. Fill. *Reversible Markov chains and random walks on graphs*. Berkeley, 2002.

A. Aleta and Y. Moreno. Multilayer Networks in a Nutshell. *Annual Review of Condensed Matter Physics*, 10(1), 2019.

O. Artime, J. J. Ramasco, and M. S. Miguel. Dynamics on networks: competition of temporal and topological correlations. *Scientific Reports*, 7:41627, 2017.

R. D. Astumian. Thermodynamics and Kinetics of a Brownian Motor. *Science*, 276 (5314):917–922, 1997.

E. Avineri. A Cumulative Prospect Theory Approach to Passengers Behavior Modeling: Waiting Time Paradox Revisited. *Journal of Intelligent Transportation Systems*, 8(4):195–204, 2004.

P. Bak, K. Christensen, L. Danon, and T. Scanlon. Unified Scaling Law for Earthquakes. *Physical Review Letters*, 88(17), 2002.

R. Bandari, S. Asur, and B. A. Huberman. The Pulse of News in Social Media: Forecasting Popularity. *CoRR*, abs/1202.0332, 2012.

P. Bao, H.-W. Shen, W. Chen, and X.-Q. Cheng. Cumulative Effect in Information Diffusion: Empirical Study on a Microblogging Network. *PLoS one ONE*, 8(10): e76027+, 2013.

A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435 (7039):207–211, 2005.

A.-L. Barabási. *Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades*. Plume, 2010.

A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

A. Bhowmick, M. Gueuning, J.-C. Delvenne, R. Lambiotte, and B. Mitra. Temporal sequence of retweets help to detect influential nodes in social networks. *IEEE Transactions on Computational Social Systems*, 2019.

A. K. Bhowmick, M. Gueuning, J.-C. Delvenne, R. Lambiotte, and B. Mitra. Temporal Pattern of (Re)tweets Reveal Cascade Migration. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 - ASONAM 17*. ACM Press, 2017.

A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. online, 2016.

S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, 2014.

M. Boguñá and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66(4), 2002.

A. Bovet and H. A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1), 2019.

S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

A. Bruns, T. Highfield, and J. Burgess. The Arab Spring and Social Media Audiences. *American Behavioral Scientist*, 57(7):871–898, 2013.

R. S. Burt. The social structure of competition. *Explorations in economic sociology*, 65:103, 1993.

A. Caprara, P. Toth, and M. Fischetti. Algorithms for the Set Covering Problem. *Annals of Operations Research*, 98(1/4):353–371, 2000.

A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti. Emergence of network features from multiplexity. *Scientific Reports*, 3(1), 2013.

S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of Internet topology using k-shell decomposition. *PNAS*, 104(27):11150–11154, 2007.

E. Çinlar. Markov renewal theory. *Advances in Applied Probability*, 1(02):123–187, 1969.

M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, pages 10–17, 2010.

D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security*, 10 (4):1–26, 2008.

D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou. Identifying influential nodes in complex networks. *Physica A*, 391(4):1777–1787, 2012.

J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web - WWW 14*. ACM Press, 2014.

J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec. Do Cascades Recur? In *Proceedings of the 25th International Conference on World Wide Web - WWW 16*. ACM Press, 2016.

F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1996.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. *Nature communications*, 6:6868, 2015.

J.-C. Delvenne, S. N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29): 12755–12760, 2010.

J.-C. Delvenne, R. Lambiotte, and L. E. C. Rocha. Diffusion on networked systems is a question of time or structure. *Nature Communications*, 6:7366, 2015.

C. Ding and K. Li. Topologically biased random walk for diffusions on multiplex networks. *Journal of Computational Science*, 2017.

L. Ding, J. Wang, and W. Wei. Method for Detecting Key Nodes who Occupy Structural Holes in Social Network sites. In *20th Pacific Asia Conference on Information Systems, PACIS 2016, Chiayi, Taiwan, June 27 - July 1, 2016*, page 174, 2016.

Y. Ding, E. Yan, A. Frazho, and J. Caverlee. PageRank for ranking authors in co-citation networks. *ASIST*, 60(11):2229–2243, 2009.

Z.-y. Ding, Y. Jia, B. Zhou, Y. Han, L. He, and J.-f. Zhang. Measuring the spreadability of users in microblogs. *Journal of Zhejiang University SCIENCE C*, 14(9):701–710, 2013.

M. D. Domenico, A. Sole, S. Gomez, and A. Arenas. Random Walks on Multiplex Networks.

M. D. Domenico, A. Sole-Ribalta, S. Gomez, and A. Arenas. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356, 2014.

M. D. Domenico, C. Granell, M. A. Porter, and A. Arenas. The physics of spreading processes in multilayer networks. *Nature Physics*, 12(10):901–906, 2016.

R. Dong, L. Li, Q. Zhang, and G. Cai. Information Diffusion on Social Media During Natural Disasters. *TCSS*, 5(1):265–276, 2018.

W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2, 2nd Edition*. John Wiley & Sons, Inc., 1971.

R. P. Feynman, R. B. Leighton, M. Sands, and E. M. Hafner. The Feynman Lectures on Physics Vol. I. *American Journal of Physics*, 33(9):750–752, 1965.

D. Figueiredo, P. Nain, B. Ribeiro, E. d. S. e Silva, and D. Towsley. Characterizing continuous time random walks on time varying graphs. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems - SIGMETRICS 12*. ACM Press, 2012.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35, 1977.

R. Gallotti and M. Barthelemy. The multilayer temporal network of public transport in Great Britain. *Scientific Data*, 2:140056, 2015.

S. Gao, J. Ma, and Z. Chen. Modeling and Predicting Retweeting Dynamics on Microblogging Platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM 15*. ACM Press, 2015.

M. Gardner. The paradox of the nontransitive dice and the elusive principle of indifference. *Scientific American*, 223:110–114, 1970.

T. D. Gauthier. Detecting trends using Spearman's rank correlation coefficient. *Environmental forensics*, 2(4):359–362, 2001.

L. Gauvin, A. Panisson, C. Cattuto, and A. Barrat. Activity clocks: spreading dynamics on temporal networks of human contact. *Scientific Reports*, 3(1), 2013.

V. Gemmetto, A. Barrat, and C. Cattuto. Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Diseases*, 14(1), 2014.

M. Génois, C. L. Vestergaard, J. Fourner, A. Panisson, I. Bonmarin, and A. Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(03):326–347, 2015.

R. Ghosh, T. Surachawala, and K. Lerman. Entropy-based Classification of 'Retweeting' Activity on Twitter. *CoRR*, abs/1106.0346, 2011.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

D. F. Gleich. PageRank Beyond the Web. *SIAM Review*, 57(3):321–363, 2015.

S. Goel, A. Anderson, J. Hofman, and D. J. Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2015.

K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002, 2008.

D. Goldenberg, A. Sela, and E. Shmueli. Timing Matters: Influence Maximization in Social Networks Through Scheduled Seeding. *TCSS*, 99:1–18, 2018.

S. Gómez, A. Díaz-Guilera, J. Gómez-Gardeñes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas. Diffusion Dynamics on Multiplex Networks. *Physical Review Letters*, 110(2), 2013.

V. Gómez, H. J. Kappen, N. Litvak, and A. Kaltenbrunner. A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, 16(5-6):645–675, 2012.

J. Gómez-Gardeñes, M. de Domenico, G. Gutiérrez, A. Arenas, and S. Gómez. Layer–layer competition in multiplex complex networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2056):20150117, 2015.

M. Gomez Rodriguez, L. Song, H. Daneshmand, and B. Schölkopf. Estimating Diffusion Networks: Recovery Conditions, Sample Complexity and Soft-thresholding Algorithm. *Journal of Machine Learning Research*, 17(90):1–29, 2016.

M. Gueuning, S. Cheng, R. Lambiotte, and J.-C. Delvenne. Rock–paper–scissors dynamics from random walks on temporal multiplex networks. *Journal of Complex Networks*, cnz027.

M. Gueuning, J.-C. Delvenne, and R. Lambiotte. Imperfect spreading on temporal networks. *The European Physical Journal B*, 88(11), 2015.

M. Gueuning, R. Lambiotte, and J.-C. Delvenne. Backtracking and Mixing Rate of Diffusion on Uncorrelated Temporal Networks. *Entropy*, 19(10):542, 2017.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu. Joint Community and Structural Hole Spanner Detection via Harmonic Modularity. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press, 2016.

X. He and D. Kempe. Robust Influence Maximization. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 885–894, New York, NY, USA, 2016. ACM.

T. Hoffmann, M. A. Porter, and R. Lambiotte. Random Walks on Stochastic Temporal Networks. In *Understanding Complex Systems*, pages 295–313. Springer Berlin Heidelberg, 2013.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.

C.-Y. Huang, Y.-H. Fu, and C.-T. Sun. Identify Influential Social Network Spreaders. In *2014 IEEE International Conference on Data Mining Workshop*. IEEE, 2014.

D.-W. Huang and Z.-G. Yu. Dynamic-Sensitive centrality of nodes in temporal networks. *Scientific reports*, 7:41454, 2017.

J. L. Iribarren and E. Moro. Branching dynamics of viral information spreading. *Physical Review E*, 84(4), 2011.

S. Jang, J. S. Lee, S. Hwang, and B. Kahng. Ashkin-Teller model and diverse opinion phase transitions on multiplex networks. *Physical Review E*, 92(2), 2015.

B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2), 2009.

Z. Jianqiang, G. Xiaolin, and T. Feng. A New Method of Identifying Influential Users in the Micro-Blog Networks. *IEEE Access*, 5:3008–3015, 2017.

H.-H. Jo, M. Karsai, J. Kertész, and K. Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.

M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2), 2011.

M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2(1), 2012.

M. Karsai, H.-H. Jo, and K. Kaski. *Bursty Human Dynamics (SpringerBriefs in Complexity)*. Springer, 2017.

M. W. Kearney. *rtweet: Collecting Twitter Data*, 2019. URL `https://cran.r-project.org/package=rtweet`. R package version 0.6.9.

D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD03*. ACM Press, 2003.

T. Kemuriyama, H. Ohta, Y. Sato, S. Maruyama, M. Tandai-Hiruma, K. Kato, and Y. Nishida. A power-law distribution of inter-spike intervals in renal sympathetic nerve activity in salt-sensitive hypertension-induced chronic heart failure. *Biosystems*, 101(2):144–147, 2010.

W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, J. Saramäki, and M. Karsai. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(03):P03005, 2012.

M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014.

S. D. Kleban and S. H. Clearwater. Hierarchical Dynamics, Interarrival Times, and Performance. In *Proceedings of the 2003 ACM/IEEE conference on Supercomputing - (SC) 03*. ACM Press, 2003.

K.-K. Kleineberg and M. Boguñá. Competition between global and local online social networks. *Scientific Reports*, 6(1), 2016.

R. Kobayashi and R. Lambiotte. TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. In *ICWSM*, 2016.

R. Lambiotte, L. Tabourier, and J.-C. Delvenne. Burstiness and spreading on temporal networks. *The European Physical Journal B*, 98(7):052307, 2013.

R. Lambiotte, V. Salnikov, and M. Rosvall. Effect of memory on the dynamics of random walks on networks. *Journal of Complex Networks*, 3(2):177–188, 2014.

G. F. Lawler. *Random Walk and the Heat Equation (Student Mathematical Library)*. American Mathematical Society, 2010.

S.-Y. Liu, A. Baronchelli, and N. Perra. Contagion dynamics in time-varying metapopulation networks. *Phys. Rev. E*, 87:032805, 2013.

T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22nd international conference on World Wide Web - WWW 13*. ACM Press, 2013.

L. Lovász. Random Walks on Graphs: A Survey. *Bolyai Society Mathematical Studies*, 2:1–46, 1993.

A. Madotto and J. Liu. Super-Spreader Identification Using Meta-Centrality. *Scientific reports*, 6:38994, 2016.

M. Magnani, B. Micenkova, and L. Rossi. Combinatorial Analysis of Multiple Networks. *arXiv preprint*, 2013.

F. D. Malliaros, M.-E. G. Rossi, and M. Vazirgiannis. Locating influential nodes in complex networks. *Scientific reports*, 6:19307, 2016.

R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158, 2008.

R. Mastrandrea, J. Fournet, and A. Barrat. Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. *PLOS ONE*, 10(9):e0136497, 2015.

N. Masuda and R. Lambiotte. *A guide to temporal networks*. World Scientific Publishing Europe Ltd, 2016.

N. Masuda and L. E. C. Rocha. A Gillespie Algorithm for Non-Markovian Stochastic Processes. *SIAM Review*, 60(1):95–115, 2018.

N. Masuda, T. Takaguchi, N. Sato, and K. Yano. Self-Exciting Point Process Modeling of Conversation Event Sequences. In *Understanding Complex Systems*, pages 245–264. Springer Berlin Heidelberg, 2013.

N. Masuda, M. A. Porter, and R. Lambiotte. Random walks and diffusion on networks. *Physics Reports*, 716-717:1–58, 2017.

S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson. The unreasonable effectiveness of tree-based theory for networks with clustering. *Physical Review E*, 83(3), 2011.

F. Morone, B. Min, L. Bo, R. Mari, and H. A. Makse. Collective influence algorithm to find influencers via optimal percolation in massively large social media. *Nature*, 6:30062, 2016.

M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

J. D. Noh and H. Rieger. Random Walks on Complex Networks. *Physical Review Letters*, 92(11), 2004.

J. G. Oliveira and A.-L. Barabási. Darwin and Einstein correspondence patterns. *Nature*, 437(7063):1251–1251, 2005.

R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and Correlation Properties of the Internet. *Physical Review Letters*, 87(25), 2001.

R. Pastor-Satorras, C. Castellano, P. V. Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3):925–979, 2015.

J. Petit, M. Gueuning, T. Carletti, B. Lauwens, and R. Lambiotte. Random walk on temporal networks with lasting edges. *Phys. Rev. E*, 98:052307, 2018.

P. Pietro, O. Tore, and C. K. M. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5):911–932, 2009.

S. Pramanik, Q. Wang, M. Danisch, S. Bandi, A. Kumar, J.-L. Guillaume, and B. Mitra. On the Role of Mentions on Tweet Virality. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2016.

F. Riquelme and P. González-Cantergiani. Measuring user influence on Twitter: A survey. *IPM*, 52(5):949–975, 2016.

L. E. C. Rocha and N. Masuda. Random walk centrality for temporal networks. *New Journal of Physics*, 16(6):063023, 2014.

L. E. C. Rocha, F. Liljeros, and P. Holme. Information dynamics shape the sexual networks of Internet-mediated prostitution. *Proceedings of the National Academy of Sciences*, 107(13):5706–5711, 2010.

L. E. C. Rocha, F. Liljeros, and P. Holme. Simulated Epidemics in an Empirical Spatiotemporal Network of 50, 185 Sexual Contacts. *PLoS Computational Biology*, 7(3):e1001109, 2011.

M. G. Rodriguez, J. Leskovec, and B. Schölkopf. Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM 13*. ACM Press, 2013.

J. Roesslein. tweepy Documentation Release 3.7.0, 2019.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4): 1118–1123, 2008.

M. Rosvall, A. V. Esquivel, A. Lancichinetti, J. D. West, and R. Lambiotte. Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, 5, 2014.

B. Ruhnau. Eigenvector-centrality—a node-centrality? *Social networks*, 22(4):357–365, 2000.

J. Saramäki and P. Holme. Exploring temporal networks with greedy walks. *The European Physical Journal B*, 88(12), 2015.

A. Scherrer, P. Borgnat, E. Fleury, J.-L. Guillaume, and C. Robardet. Description and simulation of dynamic mobility networks. *Computer Networks*, 52(15):2842–2858, 2008.

I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications*, 5:5024, 2014.

S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the national academy of sciences*, 113(36):9977–9982, 2016.

S. Seo. *A review and comparison of methods for detecting outliers in univariate data sets*. PhD thesis, University of Pittsburgh, 2006.

A. Sheikhahmadi and M. A. Nematbakhsh. Identification of multi-spreader users in social networks for viral marketing. *Journal of Information Science*, 43(3):412–423, 2017.

H. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 291–297. AAAI Press, 2014.

L. Speidel, R. Lambiotte, K. Aihara, and N. Masuda. Steady state and mean recurrence time for random walks on stochastic temporal networks. *Physical Review E*, 91(1), 2015.

E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou. Temporal Dynamics of Information Diffusion in Twitter: Modeling and Experimentation. *TCSS*, 5(1):256–264, 2018.

M. Starnini, A. Baronchelli, A. Barrat, and R. Pastor-Satorras. Random walks on temporal networks. *Physical Review E*, 85(5), 2012.

J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. V. den Broeck, C. Régis, B. Lina, and P. Vanhems. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, 6(8):e23176, 2011.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010.

M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

J. Tang, X. Tang, and J. Yuan. Influence Maximization Meets Efficiency and Effectiveness: A Hop-Based Approach. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 64–71, New York, NY, USA, 2017. ACM.

G. Tavares and A. Faisal. Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users. *PloS one*, 8(7):e65774, 2013.

E. Tonkin, H. D. Pfeiffer, and G. Tourte. Twitter, information sharing and the London riots? *Bulletin of the American Society for Information Science and Technology*, 38 (2):49–57, 2012.

F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1025–1028. ACM, 2013.

P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, and N. Voirin. Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors. *PLoS ONE*, 8(9):e73970, 2013.

A. Vazquez. Impact of memory on human dynamics. *Physica A: Statistical Mechanics and its Applications*, 373:747–752, 2007.

S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

P. Wang, T. Zhou, X.-P. Han, and B.-H. Wang. Modeling correlated human dynamics with temporal preference. *Physica A: Statistical Mechanics and its Applications*, 398:145–151, 2014.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

H. J. Wearing, P. Rohani, and M. J. Keeling. Appropriate Models for the Management of Infectious Diseases. *PLoS Medicine*, 2(7):e174, 2005.

L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting Successful Memes using Network and Community Structure. *ICWSM*, 2014.

M. S. Wheatland, P. A. Sturrock, and J. M. McTiernan. The Waiting-Time Distribution of Solar Flare Hard X-Ray Bursts. *The Astrophysical Journal*, 509(1):448–455, 1998.

B. Wilder, A. Yadav, N. Immorlica, E. Rice, and M. Tambe. Uncharted but Not Uninfluenced: Influence Maximization with an Uncertain Network. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, pages 1305–1313, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

G. Wolfsfeld, E. Segev, and T. Sheafer. Social media and the Arab Spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013.

C. Xia, S. Guha, and S. Muthukrishnan. Targeting algorithms for online social advertising markets. In *2016 EEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016.

J. Yang and S. Counts. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. *ICWSM*, 10:355–358, 2010.

P. Zhang, W. Chen, X. Sun, Y. Wang, and J. Zhang. Minimizing Seed Set Selection with Probabilistic Coverage Guarantee in a Social Network. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1306–1315, New York, NY, USA, 2014. ACM.

K. Zhao, M. Karsai, and G. Bianconi. Entropy of Dynamical Social Networks. *PLoS ONE*, 6(12):e28116, dec 2011.

Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *Proceedings of the 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1513–1522, 2015.