Application of stochastic processes to real-time bidding and diffusion processes on networks

Rémi Lemonnier

NNT : 2016SACLN068

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE À L'ECOLE NORMALE SUPÉRIEURE
PARIS-SACLAY

Ecole doctorale n°574
Ecole Doctorale de Mathématique Hadamard
Spécialité de doctorat : Mathématiques appliquées

par

M. RÉMI LEMONNIER

Applications des processus stochastiques aux enchères en temps réel et à la propagation d'information dans les réseaux sociaux

Thèse présentée et soutenue à Cachan, le 22 novembre 2016.

Composition du Jury :

M.	EMMANUEL BACRY	Ecole Polytechnique	(Président du jury)
M.	MANUEL GOMEZ-RODRIGUEZ	Max Planck Institute	(Rapporteur)
M.	MARC HOFFMANN	Université Paris Dauphine	(Rapporteur)
M.	FLORENT KRZAKALA	ENS Ulm	(Rapporteur)
M.	NICOLAS VAYATIS	ENS Paris-Saclay	(Directeur de thèse)

Remerciements

Je tiens tout d'abord à exprimer ma gratitude envers mon directeur de thèse, Nicolas Vayatis. Sa disponibilité et son accessibilité ont été des atouts extrêmement précieux tout au long des trois années écoulées, et cette thèse n'aurait pas été possible sans ses conseils.

Evidemment, je dois également beaucoup à 1000mercis, et en particulier à ses deux fondateurs, Yseulys et Thibaut. Leur goût pour la recherche, leur confiance et leur intérêt constant pour l'avancée de mes travaux ont fait de 1000mercis un formidable terreau pour cette thèse CIFRE, et ce n'est qu'un juste retour des choses que les algorithmes développés aient eu un réel impact en production. Merci également aux membres de mon « comité de pilotage » : Anne, Sandra, et Djoule, qui m'aura montré ce que signifie l'expression « management humain ».

C'est pour moi un honneur d'avoir eu pour reviewers Emmanuel Bacry, Manuel Gomez-Rodriguez, Marc Hoffman et Florent Krzakala. Leur production scientifique aura été d'une grande source d'inspiration pendant ces trois années.

Un grand merci à mes co-doctorants du CMLA, qui ont stressé avec moi devant les impitoyables deadlines et reviews des conférences, et grelotté avec moi à Montréal en Décembre en cas de succès. En particulier, merci à Kevin, mon co-auteur favori, sans qui cette thèse serait sans doute beaucoup moins fournie.

Ces trois années passées à 1000mercis ont été une expérience absolument formidable, et ce en grande partie grâce aux personnes absolument formidables qui y travaillent. Merci à Jérôme, qui aura été avec moi du début à la fin, et qui est maintenant bien plus qu'un ex-collègue. Je lui souhaite un succès fulgurant dans ses nouvelles aventures entrepreneuriales. Merci à Stéphane, qui est parti combattre la pollution en écrivant des best-sellers, à Vincent et Bastien, grands gourous du cluster Hadoop, à mes collègues data scientists Pierre, Romain, Marie, Gautier, Oana, Manon et Kevin, qui m'ont donné envie de me lever tous les matins, et à l'ensemble de l'équipe RTB, dont je suis sûr qu'ils continueront collectivement à faire de grandes choses.

Certaines amitiés sont bonnes pour la santé. C'est les cas des Skipponts et assimilés, Pierre, Casimir, Paul, Marc, Simon, Madi, Amarou et Samir, que j'ai toujours autant de plaisir à retrouver autour d'un verre qu'un banc de développé-couché. Un grand merci à

Julien, avec qui je partage depuis maintenant un certain temps mes projets les plus fous, et ne compte pas m'arrêter en si bon chemin.

Mes remerciements seraient incomplets si je ne citais pas les personnes qui m'ont soutenu inconditionnellement depuis le début. Mon père, qui m'a contaminé avec le virus de la recherche scientifique, ma mère (nekini hamlaghkem attas), ma vieille sœur et tout le reste de ma famille, de chaque côté de la méditerranée. Enfin, Karine, tu sais l'importance que tu as eu dans la réalisation de cette thèse, notre voyage ne fait que commencer.

Abstract

In this thesis, which is the result of a CIFRE collaboration with the pioneering marketing agency 1000mercis, we study two applications of stochastic processes in internet marketing. The first chapter focuses on internet user scoring for real-time bidding. This problem consists in finding the probability for a given user to perform an action of interest, called conversion, in the next few days. We show that Hawkes processes [1] are well suited for modeling such phenomena but that state-of-the-art algorithms are not applicable to the size of data sets typically involved in industrial applications. We therefore develop two new algorithms able to perform nonparametric multivariate Hawkes process inference orders of magnitude faster than previous methods. We show empirically that the first one outperforms state-of-the-art competitors, and the second one scales to very large datasets while maintaining very high prediction power. The resulting algorithms have been integrated in production and are used on everyday basis with remarkable performance in 1000mercis, where they became an important business asset. The second chapter focuses on diffusion processes on graphs, an important tool for modelizing the spread of a viral marketing operation over social networks. We derive the first theoretical bounds for the total number of nodes reached by a contagion for general graphs and diffusion dynamics. For any graph of size n, we show the existence of two distinct regimes: the sub-critical one where at most $O(\sqrt{n})$ nodes are infected, and the super-critical one where $O(n)$ nodes can be infected. We also study the behavior w.r.t. to the observation time T and reveal the existence of critical times under which a long-term super-critical diffusion process behaves sub-critically. Finally, we extend our work to different application fields, and improve state-of-the-art results in percolation and epidemiology.

Keywords: Multivariate Hawkes processes, Real-time bidding, Diffusion processes on graphs, Information cascades, Influence maximization, Viral marketing, Bond percolation, SIR Model, Large-scale nonparametric inference

Résumé

Dans cette thèse, nous étudions deux applications des processus stochastiques au marketing internet.

Le premier chapitre s'intéresse au scoring d'internautes pour les enchères en temps réel. Ce problème consiste à trouver la probabilité qu'un internaute donné réalise une action d'intérêt, appelée conversion, dans les quelques jours suivant l'affichage d'une bannière publicitaire. Nous montrons que les processus de Hawkes [1] constituent une modélisation naturelle de ce phénomène mais que les algorithmes de l'état de l'art ne sont pas applicables à la taille des données typiquement à l'œuvre dans des applications industrielles. Nous développons donc deux nouveaux algorithmes d'inférence non-paramétrique qui sont plusieurs ordres de grandeurs plus rapides que les méthodes précédentes. Nous montrons empiriquement que le premier a de meilleures performances que les compétiteurs de l'état de l'art, et que le second permet une application à des jeux de données encore plus importants sans payer un prix trop important en terme de pouvoir de prédiction. Les algorithmes qui en découlent ont été implémentés avec de très bonnes performances depuis plusieurs années à 1000mercis, l'entreprise spécialiste du marketing interactif étant le partenaire industriel de cette thèse CIFRE, où ils ont beaucoup apporté en production.

Le deuxième chapitre s'intéresse aux processus diffusifs sur les graphes qui constituent un outil important pour modéliser la propagation d'une opération de marketing viral sur les réseaux sociaux. Nous établissons les premières bornes théoriques sur le nombre total de nœuds atteint par une contagion dans le cadre de graphes et dynamiques de diffusion quelconques, et montrons l'existence de deux régimes bien distincts : le régime sous-critique où au maximum $O(\sqrt{n})$ nœuds seront infectés, où n est la taille du réseau, et le régime sur-critique ou $O(n)$ nœuds peuvent être infectés. Nous étudions également le comportement par rapport au temps d'observation T et mettons en lumière l'existence de temps critiques en-dessous desquels une diffusion, même sur-critique sur le long terme, se comporte de manière sous-critique. Enfin, nous étendons nos travaux à la percolation et l'épidémiologie, où nous améliorons les résultats existants.

Keywords: Processus de Hawkes, Real-time bidding, Processus diffusifs sur les graphes, Cascades d'information, Maximisation d'influence, Marketing viral, Bond Percolation, Modèle SIR, Inférence non-paramétrique à grande échelle

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Cette thèse est le fruit d'un contrat CIFRE (Conventions Industrielles de Formation et de Recherche) entre le laboratoire du CMLA à l'Ecole Normale Supérieure de Cachan, et 1000mercis, entreprise pionnière du marketing digital en France. Cette collaboration est en partie motivée par la profonde mutation qu'a subi le marché de la publicité en ligne ces dernières années, et notamment le changement de paradigme causé par l'apparition de l'achat *programmatique*. Alors que le modèle traditionnel de l'achat d'espace publicitaire était les transactions de gré à gré entre un *acheteur* (e.g 1000mercis) et un *publieur*, ou vendeur d'espace publicitaire (e.g lepoint.fr), les évolutions technologiques ont permis l'émergence d'*ad networks*, des places de marché en ligne permettant la mise en place d'enchères en temps réel pour chaque opportunité d'impression. Le fonctionnement simplifié d'une telle enchère est le suivant. A un instant t, l'information qu'un utilisateur u se trouve sur une page internet w est envoyée à la place de marché. Divers acheteurs reçoivent cette information et envoient en retour leur offre, notamment en fonction de leurs connaissances précédentes sur u. A $t' = t + 100ms$, chaque emplacement de w est attribué à l'acheteur ayant effectué la meilleure offre. Notons que chaque acteur enregistre en particulier l'information « u était sur w à t » (u,w,t). Comme on pouvait s'y attendre, la possibilité pour les agences de pouvoir proposer un prix différent par individu et par emplacement publicitaire leur a permis de grandement améliorer l'efficacité de leurs campagnes marketing, ce qui s'est traduit par une augmentation de la part de l'achat programmatique dans la publicité digitale qui est par exemple passée en France de 7% en 2012 à 40% en 2015 [2]. Ces nouvelles possibilités ont poussé les spécialistes du marketing interactif comme 1000mercis à mettre en place des algorithmes d'enchères à l'individu. Les travaux présentés dans cette thèse ont pour mission de répondre à ces problématiques en proposant des méthodes permettant de modéliser quantitativement le comportement des internautes, et d'utiliser ces modèles pour maximiser l'impact des campagnes de marketing programmatique. La première partie de

ce manuscrit propose une modélisation des historiques de navigation des internautes par des « processus de Hawkes » [1]. Ces processus peuvent être vus comme les analogues des séries temporelles en temps continu et permettent de quantifier précisément quel est l'impact de l'occurrence de chaque évènement de type (u, w, s) sur la probabilité qu'un évènement d'intérêt (e.g l'achat sur le site d'un client) intervienne dans un futur proche. La contribution mathématique de cette partie est le développement des premières méthodes d'inférence permettant d'utiliser ces processus pour modéliser des jeux de données de très grande taille comme ceux de 1000mercis (couramment de l'ordre de 10^{10} évènements de navigation répartis sur 10^8 internautes). Cependant, si cette modélisation des historiques de navigation des internautes en tant que réalisations indépendantes d'un même processus stochastique est souvent pertinente, elle ne permet pas de prendre en compte la façon dont les individus s'influencent les uns les autres, et donc le caractère viral des informations de marketing. La deuxième partie étudie quant à elle ces phénomènes de propagation d'information dans les graphes sociaux, notamment dans le but de pouvoir y prédire l'impact d'opérations de marketing viral. Les résultats obtenus dans cette partie sont en outre transposés avec succès dans d'autres champs d'applications, notamment l'épidémiologie et la percolation. Sans s'attarder sur les résultats techniques qui seront abordés dans le corps de la thèse, cette introduction présente les principaux résultats obtenus, ainsi que leur motivation.

1.1 Introduction à la modélisation de l'internaute par processus de Hawkes

Une campagne marketing programmatique est généralement évaluée sur sa capacité à générer un grand nombre de *conversions* (i.e achats, remplissage d'un formulaire, demande de rappel) sur le site d'un client pour un budget donné. L'attribution d'une conversion à une impression publicitaire donnée constitue un sujet de recherche à part entière [3], mais la majeure partie du temps l'annonceur et l'agence marketing se mettent d'accord sur une *fenêtre d'attribution*. Par exemple, la conversion sera attribuée à la dernière bannière publicitaire cliquée datant de moins de 30 jours, ou à défaut à la dernière bannière publicitaire affichée datant de moins de 7 jours. Dans le cas où aucune impression ne se trouve dans la fenêtre d'attribution, elle pourra être considérée comme provenant du trafic naturel du site, ou d'autres sources marketing (TV, affichage physique...). La taille des fenêtres d'attributions varie en fonction du client, et en particulier du temps caractéristique de décision d'achat du produit considéré. Le but de 1000mercis, une fois la règle d'attribution fixée, est de maximiser le nombre de conversions générées par ses campagnes marketing. Pour ce faire, il est primordial d'identifier les

individus les plus à même d'effectuer cette conversion dans le délai de la fenêtre d'attribution. Une première piste est de cibler en priorité les individus ayant déjà visité le site du client, on parle alors de *retargeting*. Une pratique courante dans l'industrie est de créer une ligne de campagne par couple degré d'engagement (i.e visite de page produit, mise au panier...) / récence de la visite (i.e <2h, 2h-24h, 24h-7j...). On observe alors selon les différentes pistes et le client des taux de conversion à 7 jours de 10 à 10 000 fois supérieurs que la moyenne de la population. Cependant, même s'il apporte des résultats corrects en pratique, ce découpage reste arbitraire et peut être largement amélioré par la mise en place d'un réel algorithme de scoring. Par ailleurs, pour la grande majorité des annonceurs, l'intérêt principal réside dans l'*acquisition*, i.e le ciblage d'individus ne s'étant jamais rendus sur leur site, et qui ont donc de fortes chances d'être des potentiels nouveaux clients. Il est donc nécessaire de répondre au problème suivant : étant donné un historique de navigation $h = (t_i^h, u_i^h)_{i\in[1...n_h]}$ constitué des n_h informations « réalisation de l'évènement u_i^h au temps t_i^h », quelle est la probabilité d'observer un évènement d'intérêt (e.g en toute généralité l'évènement d'indice 0) dans un futur proche ? C'est précisément la raison d'être des processus de Hawkes.

1.1.1 Introduction aux Processus de Hawkes

Les processus de Hawkes font partie de la famille des processus ponctuels (*point processes* en anglais), ce qui signifie qu'ils modélisent l'occurrence d'évènements en temps continu. Informellement, un processus ponctuel, aussi appelé processus de comptage, est un processus à sauts $N^t = (N_u^t)_{u=1...U}$ tel que N_u^t est égal au nombre d'évènements s'étant produits dans la dimension u entre 0 et t. Une quantité très importante des processus ponctuels est leur intensité conditionnelle stochastique $\lambda(t) = (\lambda_u(t))_{u=1...U}$ définie par

$$\lambda_u(t) = \lim_{dt\to 0} \mathbb{E}[\frac{N_u^{t+dt} - N_u^t}{dt} | \mathscr{F}_t] \tag{1.1}$$

où $\mathscr{F}_t = \sigma((N^s, \lambda(s))_{s\leq t})$ est la filtration naturelle du processus. Autrement dit, à tout instant t, sachant le passé du processus, le nombre moyen de sauts entre t et $t+dt$ dans la dimension u est $\lambda_u(t)dt$. Les processus ponctuels les plus simples sont :

- les processus de Poisson homogènes, définis par $\lambda_u(t) = \kappa$ p.s où $\kappa \in \mathbb{R}_+^n$ est un vecteur déterministe. En particulier, le nombre moyen d'évènements dans la dimension u sur tout intervalle $[s,t]$ sera $\kappa_u(t-s)$, et les évènements seront ainsi répartis uniformément sur la période d'observation.

- les processus de Poisson inhomogènes, définis par $\lambda_u(t) = \kappa(t)$ p.s où $\kappa \in C_0(\mathbb{R}_+ \to \mathbb{R}_+^n)$ est un vecteur de fonctions déterministe du temps. En particulier, le nombre

d'évènements moyen dans la dimension u sur tout intervalle $[s,t]$ sera $\int_s^t \kappa_u(x)dx$. Les évènements de dimension u seront ainsi plus fréquents sur les périodes où κ_u est grand, mais ces périodes d'intensité plus fortes sont connues avec certitude dès $t = 0$.

Un processus de Hawkes est un processus ponctuel dont l'intensité conditionnelle stochastique vérifie l'équation suivante

$$\lambda_u(t) = \left(\mu_u(t) + \sum_{v \in [1\dots d]} \sum_{t_v < t} g_{vu}(t - t_v) \right)_+ , \quad \forall u = 1, \dots, d \tag{1.2}$$

Ci-dessus, $\mu_u(t)$ représente le taux naturel d'occurrence des évènements de dimension u. Lorsque l'historique est vide, les évènements de la dimension u ont lieu comme s'ils étaient générés par un processus de Poisson inhomogène de taux $\mu_u(t)$. L'évaluation du noyau d'excitation (*triggering kernel* en anglais) $g_{uv}(t - t_v)$ donne l'évolution du taux d'occurrence dans la dimension u au temps t causée par la réalisation d'un évènement dans la dimension v au temps t_v. Suivant le signe de $g_{uv}(t - t_v)$, on parlera alors d'excitation ou d'inhibition. La modélisation des excitations mutuelles entre dimensions permet notamment de modéliser l'apparition aléatoire de périodes à très haute intensité d'évènements, ce qui n'est pas possible avec des processus de Poisson. Or ce genre de phénomènes est extrêmement courant dans de nombreux champs d'applications : citons par exemple les périodes de haute volatilité en finance, les secousses sismiques et leurs répliques en sismologie, ou les moments de haute activité neuronale en neurobiologie. Pour ce qui est du marketing internet, les internautes alternent dans la même journée entre des périodes de calme plat et des sessions de navigation où ils vont consulter de nombreuses pages. A plus grande échelle temporelle, le processus d'achat d'un internaute fait intervenir des points de contacts de plus en plus rapprochés avec la marque d'intérêt au fur et à mesure de son avancée dans le *tunnel de conversion*.

1.1.2 Contribution 1 : Inférence non-paramétrique et à grande échelle des processus de Hawkes

L'inférence des taux naturels μ_u et noyaux d'excitation g_{uv} est le plus fréquemment réalisée par maximisation de la log-vraisemblance du problème. Définissons formellement une réalisation h comme le triplet $T_h^-, T_h^+, (t_i^h, u_i^h)_{i \in [1\dots n_h]}$, où T_h^- and T_h^+ sont respectivement le début et la fin de la période d'observation. Alors la log-vraisemblance de l'ensemble de réalisations i.i.d. $\mathscr{H}$ et de $\Lambda = (M, G)$ où $M = \{\mu_u \ : \ u = 1, \dots, d\}$ et $G = \{g_{u,v} \ : \ u, v =$

$1,\dots,d\}$ est donnée par :

$$\mathscr{L}(\Lambda,\mathscr{H}) = \sum_{u=1}^{d}\sum_{h\in\mathscr{H}}\sum_{i=1}^{n_h} \ln\left(\lambda_{u_i^h}(t_i^h)\right) - \sum_{u=1}^{d}\sum_{h\in\mathscr{H}}\int_{T_h^-}^{T_h^+}\lambda_u(s)ds \tag{1.3}$$

qui se réécrit

$$\begin{aligned}\mathscr{L}(\Lambda,\mathscr{H}) &= \sum_{h\in\mathscr{H}}\sum_{i=1}^{n_h}\ln\left(\mu_{u_i^h}(t_i^h) + \sum_{j\,:\,t_j^h<t_i^h} g_{u_j^h,u_i^h}(t_i^h-t_j^h)\right)\\ &- \sum_{u=1}^{d}\sum_{h\in\mathscr{H}}\int_{T_h^-}^{T_h^+}\left(\mu_u(s) + \sum_{j=1}^{n_h} 1\left\{u_j^h=u\right\}\, g_{u,u_j}(s-t_j)\right)_+ ds\end{aligned} \tag{1.4}$$

Essentiellement, la log-vraisemblance est d'autant plus importante que le taux d'occurrence $\lambda_{u_i^h}$ était haut au temps t_i^h d'occurrence de l'évènement i pour l'historique h, et compense avec une pénalisation L^1 des grandes valeurs de λ. Une caractéristique importante de cette log-vraisemblance est le nombre d'évaluations nécessaire des $g_{u_j^h,u_i^h}$ en $O(\sum_{h\in\mathscr{H}} n_h^2)$. Si le nombre d'évènements au sein d'une même réalisation peut être important, il est souvent impossible d'effectuer ce nombre d'opérations en un temps raisonnable. Par exemple, dans le cas de la modélisation des historiques de navigation des internautes par 1000mercis, $\sum_{h\in\mathscr{H}} n_h^2$ est égal à la somme des carrés du nombre d'évènements par historique de navigation, soit de l'ordre de 10^{13}, alors que le nombre total d'évènements $\sum_{h\in\mathscr{H}} n_h$ est de l'ordre de 10^{10}. Cependant, dans certains cas particuliers, il est possible de calculer la log-vraisemblance en un temps linéaire par rapport au nombre total d'évènements. C'est notamment le cas du modèle paramétrique exponentiel, défini par $g_{uv} = \nu_{uv}\exp(-\alpha t)$ et $\mu_u = \omega_u\exp(-\alpha t)$, où chaque taux d'occurrence $\lambda_{u_i^h}(t_i^h)$ peut être calculé en temps constant grâce à la formule suivante

$$\lambda_{u_i^h}(t_i^h) = \exp\left(-\alpha(t_i^h - t_{i-1}^h)\right)\left(\lambda_{u_i^h}(t_{i-1}^h) + \nu_{u_{i-1}^h u_i^h}\right). \tag{1.5}$$

A plus haut niveau, cette formule de mise à jour vient traduire l'observation suivante : lorsque les noyaux d'excitations sont exponentiels, le processus (N_t,λ_t) est un *processus de Markov*. Cela signifie que le comportement du processus sachant son état présent est indépendant du passé, et qu'il n'est donc pas nécessaire de parcourir les évènements antérieurs pour mettre à jour l'intensité stochastique. Cependant, utiliser un noyau paramétrique exponentiel revient à faire une hypothèse très forte. En effet, la forme la plus adaptée des noyaux d'excitation dépend très fortement du champ d'application : par exemple, l'impact sur le prix des actifs d'une transaction financière donnée [4] et le processus de diffusion des vues de vidéos Youtube [5] suivent des lois puissances à queue épaisse, alors que la modélisation

des séquences ADN fait appel à des noyaux à support compact [6]. Il est donc nécessaire d'être capable d'estimer ces noyaux de manière non-paramétrique. De plus, il est hautement souhaitable d'être capable d'estimer non seulement les interactions d'excitation mutuelles, i.e $g_{uv} \geq 0$, mais également les interactions d'inhibition, qui jouent par exemple un rôle important dans la modélisation du comportement d'achat des internautes (e.g l'achat d'un objet onéreux peut pour des raisons de budget inhiber pendant plusieurs semaines le comportement d'achat sur internet d'un utilisateur). Ce problème d'estimation a fait l'objet de plusieurs travaux récents :

- Dans [7], les auteurs proposent l'algorithme MMEL basé sur une approximation low-rank des noyaux d'excitation. La log-vraisemblance est maximisée à l'aide de méthodes de type minoration-maximisation (algorithme MM). A noter que les noyaux g sont ici supposés positifs.

- Dans [8], les auteurs exhibent un contraste L^2 qui, de la même manière que l'opposé de la log-vraisemblance, fait intervenir les fonctions à estimer et l'historique des évènements du processus, et est minimal en les réels noyaux d'excitation du processus. De plus, ils proposent une méthode faisant intervenir le LASSO permettant notamment d'utiliser un dictionnaire de fonctions de bases de taille importante et d'en sélectionner les quelques meilleurs éléments de façon parcimonieuse. Des bornes théoriques d'approximation sont obtenues rigoureusement, et permettent de quantifier la décomposition biais/variance de l'erreur d'approximation.

- Dans [9, 10], les auteurs utilisent le fait que la matrice des noyaux d'excitation et les caractéristiques de second ordre du processus sont liés par une équation de Wiener-Hopf. Leur procédure se décompose en deux parties : l'estimation empirique des mesures de covariance, et la résolution de l'équation de Wiener-Hopf afin d'en déduire les noyaux d'excitation.

Cependant, chacune de ces méthodes souffre de la complexité quadratique en $O(\sum_{h \in \mathscr{H}} n_h^2)$, ce qui les rend inapplicables aux jeux de données de très grande taille.

Dans la section 2.3, nous introduisons l'algorithme MEMIP (pour *Markovian Estimation of Mutually Interacting Processes*) qui est à notre connaissance la première méthode d'estimation non-paramétrique des noyaux d'excitation et taux naturels d'occurrence des processus de Hawkes qui soit linéaire en le nombre total d'évènements considérés. Plus spécifiquement, cet algorithme s'appuie sur l'approximation des fonctions à apprendre par

des polynômes d'exponentielles :

$$\forall t \in [0,T], \quad \widehat{\mu}^K(t) = \sum_{k=0}^{K} X^K_{u0,k} \exp(-k\alpha t) \quad \text{et} \quad \widehat{g}^K_{uv}(t) = \sum_{k=0}^{K} X^K_{uv,k} \exp(-k\alpha t) \ .$$

Nous montrons que la famille de fonctions obtenue est suffisamment riche pour contenir des éléments arbitrairement proches des vrais noyaux d'excitation, et exhibons un processus de Markov qui permet le calcul de tous les taux d'occurrences en un seul passage sur le jeu de données. Une fois le modèle posé, la log-vraisemblance est concave et peut être maximisée par de multiples méthodes, mais nous présentons dans la suite une procédure basée sur la méthode de Newton qui exploite la *self-concordance* de la fonction objectif. Cette caractéristique particulière nous permet d'obtenir une borne sur le nombre total d'itérations à effectuer. Nous montrons ensuite que la méthode obtenue permet de modéliser à la fois plus efficacement et avec de meilleurs résultats divers jeux de données, aussi bien simulés que réels, ainsi que d'apporter de réels incréments de performance en pratique pour 1000mercis.

Si le travail de la section 2.3 permet d'inférer des processus de Hawkes sur des jeux beaucoup plus grands que l'état de l'art, la dépendance en le nombre de dimensions d reste toujours quadratique du simple fait de la nécessité d'apprendre d^2 noyaux d'excitations et d taux naturels d'occurrence. La complexité totale de l'algorithme est de l'ordre de $O(d^2 \sum_h n_h)$ dès que l'on dispose de suffisamment d'évènements pour que l'inférence du processus ait un sens (*i.e* $\sum_h n_h > d^2$). Dans les cas d'applications où d est très grand, la complexité algorithmique devient extrêmement importante, et le problème est mal conditionné à cause du trop grand nombre de fonctions non-paramétriques à apprendre. C'est par exemple le cas du problème de scoring d'internautes à 1000mercis où les différentes dimensions du processus sont les $d = 9.9 \times 10^8$ URLs distinctes visitées par au moins un internaute sur les trois derniers mois. Il est donc nécessaire de trouver un moyen de généraliser d'une dimension à l'autre. Une première solution est d'effectuer un clustering des dimensions a priori sur des critères extérieurs, par exemple la structure sémantique des URLs dans notre exemple d'application. Si cl est la fonction attribuant à chaque URL son cluster $c \in [1...C]$, le processus de Hawkes considéré est alors $(N'_c(t) = \sum_{u=1}^{d} \mathbb{1}_{cl(u)=c} N_u(t))_{c \in [1...C]}$. Cependant, rien ne garantit que ce clustering a priori soit réellement adapté à la modélisation par processus de Hawkes du jeu de données d'intérêt. Intuitivement, dans le cas de 1000mercis, si l'évènement à expliquer est par exemple la souscription d'un crédit à la consommation, on aimerait considérer comme proches la visite d'un site proposant des voitures et celle d'une agence de voyages, car elles traduisent toutes deux un projet onéreux pour l'internaute. Si la conversion à expliquer est la souscription à une assurance automobile, regrouper ces deux sites n'est plus du tout pertinent.

1.1.3 Contribution 2: Low-Rank Hawkes Processes

Dans la section 2.4, nous introduisons l'algorithme LRHP (pour *Low-Rank Hawkes Processes*), dont l'objectif est à la fois de répondre à ce problème de clustering optimal des dimensions et de réduire encore drastiquement la complexité de l'inférence pour les processus de Hawkes. Ce modèle se base sur la décomposition de rang faible de la matrice des noyaux d'excitations, à savoir :

$$\begin{aligned} \mu_u(t) &= \textstyle\sum_{i=1}^{r} P_{ui}\, \tilde{\mu}_i(t), \\ g_{vu}(t) &= \textstyle\sum_{i,j=1}^{r} P_{ui} P_{vj}\, \tilde{g}_{ji}(t), \end{aligned} \tag{1.6}$$

En choisissant $r \ll d$, cette approximation de rang faible a deux objectifs: i) imposer de la régularité sur les fonctions à apprendre en introduisant des contraintes sur les paramètres, et ii) réduire le nombre de paramètres. En particulier, seulement r taux naturels d'occurrence et r^2 noyaux d'excitation doivent être estimés, au lieu de d et d^2 précédemment. Le prix à payer est seulement linéaire en d : il s'agit de l'inférence des $d \times r$ éléments de la matrice P. L'inférence de la log-vraisemblance approximée est ensuite réalisée en combinant des méthodes de minoration-majoration et les algorithmes dérivés dans la section 2.3. La complexité de l'algorithme obtenu est seulement $O(d \sum_{h \in \mathscr{H}} n_h)$, ce qui lui permet d'être appliqué à des jeux de données bien plus larges que les algorithmes de la littérature. Nous montrons dans la section expérimentale que le prix à payer est relativement faible en terme de performances de prédiction, aussi bien sur des jeux de données simulés que sur des jeux de données réels.

1.2 Introduction à la propagation d'information sur les graphes

L'émergence des réseaux sociaux a eu un impact considérable sur la propagation des idées et informations. Pour les spécialistes du marketing interactif comme 1000mercis, ces nouveaux réseaux de diffusion et de communication sont devenus le terrain favori de mise en place d'opérations de *marketing viral*. A l'inverse des publicités traditionnelles, les opérations virales, souvent présentées sous forme de jeux-concours, n'ont pas nécessairement pour but de provoquer un achat immédiat de l'internaute visé, mais plutôt de l'inciter à diffuser un lien de participation à son cercle social. Les dynamiques de propagation qui en résultent sont alors de même nature que celles observées en épidémiologie [11]. Plus généralement, les dynamiques sociales ont forcé les entreprises à ré-imaginer leurs clients non comme une masse d'agents économiques isolés, mais comme des réseaux de clients fortement reliés entre eux [12]. Cependant, la compréhension des phénomènes de diffusion sur les graphes dépasse

de très loin le seul cadre du marketing sur les réseaux sociaux. Les applications possibles vont de l'étude de propagation de pandémies par les réseaux de transports aériens à celle des virus sur les réseaux informatiques, en passant par la propagation de rumeurs au sein d'un cercle social. Enfin, l'étude de ces dynamiques de propagation est très liée au domaine de la *percolation*, qui étudie la taille des composantes connexes des graphes aléatoires. Plus formellement, étant donné un graphe à n noeuds $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ et un ensemble de n_0 *influenceurs* initiaux (e.g. conscients d'une information, infectés par une maladie, ou premiers adopteurs d'un produit), la dynamique des infections virales au niveau microscopique est habituellement décrite par le modèle des *cascades d'information*, qui se décline en deux versions :

- Les cascades d'information en temps discret. En $t = 0$, seuls les influenceurs sont infectés. Etant donné une matrice $\mathscr{P} = (p_{ij})_{ij} \in [0,1]^{n \times n}$, chaque nœud i ayant reçu le virus au temps t le transmet en temps $t+1$ sur l'arête $(i, j) \in \mathscr{E}$ avec probabilité p_{ij}. Le nœud i ne peut pas infecter ses voisins aux temps suivants. Le processus continue jusqu'à ce que plus aucune infection ne soit possible.

- Les cascades d'information en temps continu. En $t = 0$, seuls les influenceurs sont infectés. Etant donné une matrice $\mathscr{F} = (f_{ij})_{ij}$ de fonctions positives et intégrables, chaque nœud i ayant reçu le virus au temps t le transmet au temps $s > t$ sur l'arête $(i, j) \in \mathscr{E}$ avec le taux d'occurrence stochastique $f_{ij}(s-t)$. Ce processus s'arrête à un temps déterministe $T > 0$. Ces modèles proposent une modélisation plus réaliste du comportement temporel d'une infection, mais leur configuration finale est équivalente à celle des modèles en temps discret si $T = \infty$.

Ces problématiques ont donné lieu à un large corpus de publications. Dans la communauté du machine learning, l'angle adopté a souvent été celui de la *maximisation d'influence* [13]. Ce problème consiste à sélectionner l'ensemble A de nœuds réseau à infecter à $t = 0$ afin de maximiser en espérance l' *influence* de A à un temps $T > 0$, i.e. le nombre total de nœuds infectés en T par l'infection provenant de A:

$$\sigma(A) = \sum_{v \in \mathscr{V}} \mathbb{P}(v \text{ est infecté par l'épidémie } |A). \tag{1.7}$$

Kempe, Kleinberg and Tardos ([14]) ont montré que, même dans le cas $T = \infty$, ce problème est NP-complet. Cependant, ils montrent également que l'algorithme glouton qui consiste à ajouter itérativement à l'ensemble d'influenceurs le nœud qui maximise l'incrément d'influence (estimé par simulations Monte-Carlo) conduit à une influence d'au minimum $(1 - \frac{1}{e})\sigma_k^{max}$ où σ_k^{max} est l'influence du meilleur ensemble de k influenceurs. Depuis, plusieurs

techniques ont été proposées pour améliorer la complexité algorithmique de l'algorithme ([15–18]). A noter que l'application en pratique de ces différents algorithmes nécessite l'inférence au préalable des caractéristiques p_{ij} de diffusion du processus, qui peut par exemple être effectuée par maximisation de log-vraisemblance [19, 20]. Cependant, à notre connaissance, peu de travaux se sont attaqués au problème de maximisation d'influence du point de vue théorique. Les contributions principales du chapitre 2 consistent en la dérivation de bornes théoriques en formules explicites sur l'influence maximale dans un graphe. Ces nouveaux résultats sont ensuite étendus à d'autres champs d'application, comme l'épidémiologie et la percolation.

1.2.1 Contribution 3: bornes exactes sur l'influence en temps infini

Les résultats qui suivent mettent en évidence que l'influence en temps infini dépend fortement d'une notion nouvelle : la *matrice des risques* (hazard matrix en anglais), quantité caractéristique du réseau, que nous définissons comme suit.

Definition 1. *Etant donné un graphe $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, équipé de probabilités de transmissions p_{ij}, la* matrice des risques *est $\mathscr{H}$ est la matrice $n \times n$ dont les coefficients sont donnés par:*

$$\mathscr{H}_{ij} = \begin{cases} -\ln(1-p_{ij}) & si\ (i,j) \in \mathscr{E} \\ 0 & sinon \end{cases} . \tag{1.8}$$

Cette matrice constitue essentiellement une transformation de la matrice des proabilités des transmissions par la bijection croissante $x \to -\ln(1-x)$, et en est d'ailleurs très proche si $\max_{i,j} p_{ij} \ll 1$. Nous introduisons également une modification de la matrice des risques pour laquelle les éléments des colonnes correspondant aux influenceurs $A \subset \mathscr{V}$ sont remplacés par des zéros:

$$\mathscr{H}(A)_{ij} = \mathbb{1}_{\{j \notin A\}} \mathscr{H}_{ij}. \tag{1.9}$$

Rappelons que pour toute matrice carrée M, son rayon spectral $\rho(M)$ est défini par $\rho(M) = \max_i(|\lambda_i|)$ où $\lambda_1, ..., \lambda_n$ sont les valeurs propres de M. Notre premier résultat principal s'applique à tout ensemble $A \subset \mathscr{V}$ d'influenceurs de cardinal n_0. Dans le cadre de la maximisation de la propagation d'une opération de marketing viral, ce résultat correspond au meilleur cas où les efforts publicitaires ont été concentrés sur les individus les plus influents.

Proposition 1. *Soit $\rho_c(A) = \rho(\frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2})$. Alors, pour tout A tel que $|A| = n_0 < n$:*

$$\sigma(A) \leq n_0 + \gamma_1(n - n_0), \tag{1.10}$$

où γ_1 est la plus petite solution dans $[0,1]$ de l'équation:

$$\gamma_1 - 1 + \exp\left(-\rho_c(A)\gamma_1 - \frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}\right) = 0. \tag{1.11}$$

Le corollaire suivant met en lumière l'existence de deux comportements bien distincts en fonction de la valeur de $\rho_c(A)$.

Corollary 1. *Sous les mêmes hypothèses que la proposition précédente:*

- *si* $\rho_c(A) < 1$, $\quad \sigma(A) \leq n_0 + \sqrt{\frac{\rho_c(A)}{1-\rho_c(A)}}\sqrt{n_0(n-n_0)},$
- *si* $\rho_c(A) \geq 1$, $\quad \sigma(A) \leq n - (n-n_0)\exp\left(-\rho_c(A) - \frac{2\rho_c(A)}{\sqrt{4n/n_0 - 3} - 1}\right).$

Ainsi, si $\rho_c(A) < 1$, le nombre de nœuds infectés sera au maximum de l'ordre de $O(\sqrt{n})$, alors que si $\rho_c(A) \geq 1$, il peut être en $O(n)$. Le deuxième résultat principal traite le cas où A est choisi uniformément parmi l'ensemble de n_0 nœuds parmi n (noté $\mathscr{P}_{n_0}(\mathscr{V})$). Dans le cadre de la maximisation de la propagation d'une opération de marketing viral, ce résultat correspond au cas où, dans l'absence d'information privilégiée, les efforts publicitaires sont répartis uniformément sur la population (campagnes dite d'*acquisition pure*).

Proposition 2. *Soit $\rho_c = \rho(\frac{\mathscr{H}+\mathscr{H}^\top}{2})$. Si l'ensemble d'influenceurs A est tiré d'une distribution uniforme sur $\mathscr{P}_{n_0}(\mathscr{V})$, alors, en notant $\sigma_{uniform}$ l'espérance du nombre de nœuds atteint par une infection partant de A:*

$$\sigma_{uniform} \leq n_0 + \gamma_2(n-n_0), \tag{1.12}$$

où γ_2 est l'unique solution dans $[0,1]$ de l'équation suivante :

$$\gamma_2 - 1 + \exp\left(-\rho_c\gamma_2 - \frac{\rho_c n_0}{n-n_0}\right) = 0. \tag{1.13}$$

Le corollaire suivant nous permet une fois encore de séparer deux comportements bien distincts en fonction de la valeur de $\rho_c(A)$.

Corollary 2. *Sous les mêmes hypothèses que la proposition précédente:*

- *si* $\rho_c < 1$, $\quad \sigma_{uniform} \leq \frac{n_0}{1-\rho_c},$
- *si* $\rho_c \geq 1$, $\quad \sigma_{uniform} \leq n - (n-n_0)\exp\left(-\frac{\rho_c}{1-\frac{n_0}{n}}\right).$

En particulier, on observe que quand $\rho_c < 1$, $\sigma_{\text{uniform}} = O(1)$, ce qui signifie en pratique que l'infection reste contenue à un très petit nombre d'individus. Les résultats expérimentaux de la section 3.1.6 montrent que ces résultats sont exacts pour une large famille de graphes, et que notamment l'influence des graphes explose effectivement pour $\rho_c = 1$ dans la quasi-totalité des cas. Les implications sont très importantes pour l'analyse des phénomènes viraux sur les réseaux sociaux. Notamment, l'inférence des probabilités de transmissions p_{ij} au début d'une opération virale et l'exploitation de ces résultats permets de se rendre compte au bout de quelques jours du potentiel de la campagne publicitaire. Si $\rho_c \ll 1$, il y a très peu de chances que le contenu devienne viral, et il est nécessaire de retravailler la promotion afin d'augmenter le taux de partages. Inversement, si ρ_c est proche de 1, voire supérieur, la campagne a un très grand potentiel et les investissement publicitaires peuvent être augmentés afin de maximiser ses chances de devenir un "buzz". Nous présentons dans les paragraphes suivant les résultats obtenus en tant que corollaire de nos résultats principaux dans d'autres champs d'application.

1.2.2 Contribution 4: nouveaux résultats en percolation

La percolation est le domaine d'étude de la distribution de la taille des composantes connexes des graphes aléatoires non dirigés. Dans le cas de la *percolation sur les arêtes*, l'ensemble des nœuds $\mathscr{V}$ est déterministe mais les arêtes $\mathscr{E}$ sont tirées aléatoirement. Une question centrale est notamment l'estimation de la taille de la plus grande composante connexe d'un graphe aléatoire. En percolation, le cas le plus simple est celui des graphes d'Erdös-Rényi $\mathscr{G}(n,p)$ où les arêtes d'un graphe complet de n nœuds sont retirées avec une probabilité $1-p$ identique pour toutes les arêtes. Dans ce cas, Erdös-Rényi [21] ont montré dans leur article fondateur que la transition de phase (i.e. la probabilité p à partir de laquelle la taille de la composante connexe grandit linéairement avec n pour n suffisamment grand) est obtenue pour $p = \frac{1}{n}$. Depuis, ces résultats ont été améliorés par Bollobas [22] et Lucszak [23] dans le cas $pn = O(1)$. Cependant, peu de résultats existent pour le cas plus général des graphes inhomogènes, où chaque arête (i,j) est retirée indépendamment avec probabilité $1-p_{ij}$. Des avancées récentes dans des cas particuliers sont l'œuvre de Janson, Bollobas and Riordan [24] quand $|\mathscr{E}| = O(n)$ et de Bollobas, Borgs, Chayes et Riordan [25] quand $|\mathscr{E}| = O(n^2)$. Dans ces articles, les auteurs dérivent de nombreux résultats asymptotiques (i.e. valides pour $n \to \infty$) tels que les valeurs des seuils de percolation et des bornes supérieures sur les tailles des plus grandes composantes connexes. Nous montrons dans la section 3.1.4 que nos résultats obtenus sur les cascades d'information permettent en fait de dériver ce qui constitue à notre connaissance les premiers résultats dans le cas général des graphes inhomogènes. L'intuition est la suivante : pour un processus de diffusion de type cascade d'informations

à temps discret, les variables p_{ij} peuvent être toutes tirées en $t = 0$, avant même que le processus d'infection commence. Dans ce cas, l'influence d'un nœud i est égale au cardinal de l'ensemble accessible (*reachable set* en anglais) à partir de i dans le graphe dirigé tel que $(i, j) \in \mathscr{E}$ si et seulement si $p_{ij} = 1$. En réalité, nous montrons que les résultats obtenus pour le problème particulier de l'estimation d'influence des cascades d'information sont encore vrais pour le problème plus général de l'estimation de taille des *ensembles atteignables* vérifiant certaines conditions, dites conditions *LPC*, qui contient notamment la percolation sur les arêtes.

1.2.3 Contribution 5: nouveaux résultats en épidémiologie

Historiquement, l'épidémiologie a été un des premiers champs d'application des modèles de diffusion, tant macroscopiques que microscopiques. L'un des plus utilisés est le modèle *Susceptible-Infected-Removed* (SIR), introduit dès 1932 [26]. Etant donné un graphe $\mathscr{G}$, un taux d'infection β et un taux de rétablissement δ, une épidémie SIR est définie par le processus stochastique $(S_t, I_t, R_t)_{t\geq 0}$ où S_t , I_t et R_t encodent l'état du réseau au temps t de sorte que $S_t^i = 1$ (resp. $I_t^i = 1$; $R_t^i = 1$) si et seulement si le nœud i est susceptible (resp. infecté ; rétabli) au temps t. A $t = 0$, une partie des nœuds est infecté (les "patients zéros") et le reste du graphe est susceptible. Chaque nœud infecté transmet ensuite l'infection selon toutes ses arêtes sortantes à un taux stochastique $\beta > 0$ (i.e. si $S_t^i = 1$ et i a k voisins infectés, alors $\lim_{dt\to 0} \mathbb{E}[I_{t+dt}^i]/dt = k\beta dt$) et se rétablit à un taux $\delta > 0$ (i.e. si $I_t^i = 1$, alors $\lim_{dt\to 0} \mathbb{E}[R_{t+dt}^i]/dt = \delta dt$). Dans le cadre d'un graphe $\mathscr{G}$ quelconque, les travaux de Draief, Ganesh et Massoulié [27] constituent à notre connaissance les premiers résultats permettant de borner le nombre de nœuds infectés en temps long. Dans le cas d'une pandémie, disposer de telles bornes a un réel intérêt pratique afin de savoir si l'infection sera ou non cantonnée à une zone géographique limitée. Nous montrons dans la section 3.1.4 que nos bornes sont aussi applicables pour le modèle SIR et généralisent les résultats de [27] en étant à la fois plus précises et moins restrictives pour ce qui est des hypothèses. Nos bornes sont aussi généralisables au cas de fonctions d'incubation plus réalistes, comme des lois log-normales [28].

1.2.4 Contribution 6: bornes sur l'influence à horizon fini $T > 0$

Si les résultats obtenus en temps infini permettent déjà d'avoir une estimation assez précise du potentiel de viralité d'une diffusion, les applications pratiques font le plus souvent intervenir un horizon $T > 0$ fini. Par exemple, en marketing, une des formes les plus répandues d'opération de collecte d'e-mails se base sur la mise en place d'un tirage au sort par lequel

chaque participant a d'autant plus de chances d'être récompensé qu'il a invité de membres de son cercle social à participer. L'objectif est alors de maximiser le nombre de souscriptions à la date du tirage au sort, les probabilités de transmissions étant quasi-nulle après cette date. Dans le cadre des cascades d'information à temps continu, l'inférence du processus de diffusion [29, 30] et des heuristiques de maximisation d'influence [31, 32] ont été développés avec succès. Cependant, en l'absence de bornes théoriques suffisamment proches de l'influence maximale dans un graphe, il était très difficile d'évaluer ces algorithmes. Il est également hautement souhaitable de disposer de formules explicites d'estimation ne nécessitant pas de simulations Monte-Carlo, trop couteuses en temps de calcul. Nous développons donc dans la section 3.2 les premières bornes théoriques à horizon fini dans le cadre des cascades d'informations à temps continu. Plus spécifiquement, nous introduisons la version en temps continu de la matrice de risques de la section 3.1, appelée *matrice de risques de Laplace*.

Definition 2. *Soit $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ un graphe dirigé, et p_{ij} des probabilités de transmissions intégrables telles que $\int_0^{+\infty} p_{ij}(t)dt < 1$.Pour $s \geq 0$, la* matrice laplacienne des risques $\mathscr{H}(s)$ *est la matrice $n \times n$ dont les coefficients sont*

$$\mathscr{H}_{ij}(s) = \begin{cases} -\hat{p}_{ij}(s) \left(\int_0^{+\infty} p_{ij}(t)dt\right)^{-1} \ln\left(1 - \int_0^{+\infty} p_{ij}(t)dt\right) & si\ (i,j) \in \mathscr{E} \\ 0 & sinon \end{cases} . \tag{1.14}$$

où $\hat{p}_{ij}(s)$ désigne la transformée de Laplace de p_{ij} définie pour tout $s \geq 0$ par $\hat{p}_{ij}(s) = \int_0^{+\infty} p_{ij}(t)e^{-st}dt$.
Le résultat suivant souligne le rôle crucial joué par le rayon spectral de cette matrice dans l'influence à horizon fini $T > 0$, pour tout ensemble d'influenceurs A de cardinal $|A| = n_0$.

Proposition 3. *Soit $\rho(s) = \rho(\frac{\mathscr{H}(s)+\mathscr{H}(s)^\top}{2})$. Pour tout ensemble d'influenceurs initiaux A tel que $|A| = n_0 < n$, l'espérance $\sigma_A(T)$ du nombre de noeuds atteint par la contagion au temps T vérifie :*

$$\sigma_A(T) \leq n_0 + (n - n_0) \min_{s \geq 0} \gamma(s) e^{sT}. \tag{1.15}$$

où $\gamma(s)$ est la plus petite solution dans $[0,1]$ de l'équation suivante :

$$\gamma(s) - 1 + \exp\left(-\rho(s)\gamma(s) - \frac{\rho(s)n_0}{\gamma(s)(n-n_0)}\right) = 0. \tag{1.16}$$

Ce résultat est une généralisation de la version en temps infini de la Sec. 3.1 qui est retrouvée en prenant $s = 0$. Si la valeur optimale de s doit en général être calculée numériquement, le résultat suivant montre que, même quand $\rho(0) > 1$, il est toujours possible d'exploiter les résultats sous-critiques à un coût croissant exponentiellement avec T.

Corollary 3. *Sous les mêmes hypothèses que la proposition précédente :*

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \min_{\{s\geq 0 | \rho(s)<1\}} \left(\sqrt{\frac{\rho(s)}{1-\rho(s)}} e^{sT} \right), \tag{1.17}$$

Ces résultats nous permettent également de mettre en lumière des *temps critiques* en-dessous desquels un processus, même sur-critique, reste contrôlé. Ces résultats ont notamment un grand intérêt en épidémiologie, où savoir qu'une épidémie risque d'infecter une partie non-négligeable de la population est bien sûr très important, mais soulève une question cruciale : combien de temps ont les pouvoirs publics pour réagir ? Formellement, nous définissons la notion de temps critique à partir de limites de contagions sur des réseaux.

Theorem 1. *Soit $(\mathcal{G}_n)_{n\in\mathbb{N}}$ une suite de réseaux de taille n, et $(p^n_{ij})_{n\in\mathbb{N}}$ les fonctions de probabilités de transmissions suivant les arêtes de $\mathcal{G}_n$. Soit $\sigma_n(t)$ l'influence maximale dans $\mathcal{G}_n$ au temps t pour les ensembles d'influenceurs de taille 1. Alors il existe un* temps critique *$T^c \in \mathbb{R}_+ \cup \{+\infty\}$ tel que, pour toute suite de temps $(T_n)_{n\in\mathbb{N}}$:*

- *Si $\limsup_{n\to+\infty} T_n < T^c$, alors $\sigma_n(T_n) = o(n)$,*
- *Si $\sigma_n(T_n) = o(n)$, alors $\liminf_{n\to+\infty} T_n \leq T^c$.*

De plus, un tel temps critique est unique.

Les bornes sur l'influence en temps T nous permettent alors de dériver le corollaire suivant.

Corollary 4. *Supposons que $\forall n \geq 0, \rho_n(0) \geq 1$ et $\lim_{n\to+\infty} \frac{\rho_n^{-1}(1-\frac{1}{\ln n})}{\rho_n^{-1}(1)} = 1$. Si la suite $(T_n)_{n\in\mathbb{N}}$ vérifie*

$$\limsup_{n\to+\infty} \frac{2\rho_n^{-1}(1)T_n}{\ln n} < 1. \tag{1.18}$$

alors

$$\sigma_A(T_n) = o(n). \tag{1.19}$$

Autrement dit, le régime de contagion est *sous-critique* avant $\frac{\ln n}{2\rho_n^{-1}(1)}$ et

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\rho_n^{-1}(1)}. \tag{1.20}$$

Nous montrons dans la suite de la section 3.2 que ces résultats nous permettent d'obtenir de nouveaux résultats en épidémiologie. De plus, la section expérimentale démontre empiriquement l'exactitude des bornes au début et à la fin du processus de diffusion.

Chapter 2

Fast estimation in multivariate Hawkes processes for large scale internet user scoring

This chapter heavily relies on :

- a paper written with Nicolas Vayatis [33] published in the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery* (ECML-PKDD) 2014.
- a paper written with Kevin Scaman, Argyris Kalogeratos and Nicolas Vayatis [34] (preprint), submitted to the conference *Advances in Neural Information Processing Systems* (NIPS) 2016.

2.1 Introduction

In many real-world phenomena, such as product adoption or information sharing, events exhibit a *mutually-interacting* behavior, in the sense that the occurrence of an event will impact the occurrence rate of others events. In finance, arrivals of buying and selling orders for different stocks convey information about macroscopic market tendencies. In the study of information propagation, users of a social network share information from one to another, leading to *information cascades* spreading throughout the social graph. Closer to our industrial problem of interest, one of the industrial goals of this PhD was to determine if the purchasing behavior of a client of an online shopping website could be, to a large extent, predicted by his past navigation history on other websites.

Over the past few years, the study of point processes gained attention as the acquisition of such datasets by companies and research laboratories became increasing simple. However, the traditional models for time series analysis, such as discrete-time auto-regressive models, do not apply in this context due to fact that events happen in a continuous way. Multivariate Hawkes processes (MHP) [35, 36] have emerged in several fields as the gold standard to deal with such data, e.g. earthquake prediction [37, 38], biology [39], financial market analysis [40–43], social interactions studies [5], crime prediction [44] and genome analysis [6]. For MHP, an event of type u (e.g. the visit of a product's website) occurring at time t, will modify the conditional rate of occurrence of events of type v at time $s \geq t$ (e.g. purchases of this product in the future) by a rate $g_{uv}(s-t)$.

Multivariate Hawkes processes are fairly well-known from a probabilistic point of view: their Poisson cluster representation was outlined by the seminal paper of Hawkes and Oakes [1], stability conditions and sample path large deviations principles were derived in a series of papers by Brémaud and Massoulié (see e.g [45]). In the unidimensional case, Ogata [46] showed that the log-likelihood estimator enjoys usual convergence properties under mild regularity conditions. However, in practical applications, estimation of the triggering kernels g_{uv} has always been a difficult task. First, because Hawkes log-likelihood contains the logarithm of the weighted sum of triggering kernels, most of the aforementioned papers made the choice of fixing triggering kernels up to a normalization factor in order to ensure concavity, that is $g_{uv} = c_{uv} \cdot g$. Secondly, when computational efficiency is an issue, the dependency of the stochastic rate at a given time on all the past occurrences implies quadratic complexity in the number of occurrences for tasks like log-likelihood computation.

This issue has often been tackled by choosing memoryless exponential triggering kernels, but the actual dynamics of kernels strongly depends on the field of application: price impacts of a given trade [4] and process of views of Youtube videos [5] were shown to be better described by slowly decaying power-law kernels whereas for DNA sequence modelization [6] kernels are known to have bounded support. Thus, it is highly desirable to estimate triggering kernels in a data-driven way instead of assuming a given parametric form. Nonparametric estimation has been successfully addressed for unidimensional [6, 47] and symmetric bidimensional [4] Hawkes processes. In the case where triggering kernels are known to sparsely decompose over a dictionary of basis functions of bounded support (e.g for neuron spikes interactions), a LASSO-based algorithm with provable guarantees was derived in [8].

More recently, combining majorization-minimization techniques with resolution of a Euler-Lagrange equation, Zhou, Zha and Song [7] proposed, the first nonparametric learning algorithm for general multivariate Hawkes processes, named MMEL. But although this work

has constituted a significant improvement over existing parametric methods and will be the main competitor baseline used in this section, it still relies on several assumptions. First, interactions between events are assumed to be "mutually-exciting", i.e. $g_{uu'}$ are non-negative for all u, u'. We nevertheless argue that in real-world settings, there is no reason to think that interactions between events are only mutually-exciting. Secondly, the background rates μ_u are assumed to be constant. While this is a common assumption for multivariate Hawkes processes, it was shown by [48] that estimating $\mu_u(t)$ from the data could lead to significant improvement.

However, the main reason that renders intractable the application of this work to the internet user scoring problem encountered at 1000mercis is computational. Indeed, the task consists in scoring $H = 4.1 \times 10^8$ internet users (represented by their cookie) whose navigation histories involve in total $d = 9.9 \times 10^8$ URLs. The total number of events in their past three months navigation history is $\sum_{h=1}^{H} n_h = 9.0 \times 10^9$, whereas the sum of the number of events squared is $\sum_{h=1}^{H} n_h^2 = 5.7 \times 10^{12}$. Thus, the computational complexity of [7] in $O(d^2 \sum_h n_h^2)$ cannot be afforded.

To address these different issues, we construct two novel algorithms. In Sec. 2.3, we introduce MEMIP (Markovian Estimation of Mutually Interacting Processes), an algorithm based on polynomial approximation of a mapping of the triggering kernels to $[0, 1]$. Our method does not assume non-negativity on triggering kernels and is able to estimate time-dependent background rate on a data-driven way. Moreover, by constructing a markovian and linear estimator, we carry the more appealing properties of the most widely used parametric setting, where triggering kernels are fixed to exponentials up to a normalization factor : concavity of the log-likelihood that ensures global convergence of the estimator, and log-likelihood calculation in a single pass through the data achieving a complexity of $O(d^2 \sum_{h=1}^{H} n_h)$. While giving a concave formulation of the exact log-likelihood that can be maximized by multiple optimization techniques, we propose an algorithm based on maximization of a self-concordant approximation that is shown to outperform state-of-the-art methods on both simulated and real-world open data sets. Concerning the application on 1000mercis datasets, the still prohibitive complexity can be further reduced by 1) computing only the triggering kernels directly impacting the stochastic rate of the event of interest and 2) performing a clustering on the URLs in $r \ll d$ classes. Overall, these two tricks reduce the number of operations to a tractable $O(r \sum_{h=1}^{H} n_h)$. We show in Sec. 2.3.5 that the resulting algorithm is very good at selecting the best internet users and leads to a greatly improved efficiency in real-world marketing campaigns.

In Sec. 2.4, we go one step further and propose a way to reduce the inference complexity of any multivariate Hawkes process to $O(d \sum_{h=1}^{H} n_h)$ by introducing *Low-Rank Hawkes Pro-*

cesses (LRHP), a model for structured point processes relying on a *low-rank decomposition of the triggering kernel* that aim to learn representative patterns of interaction between event types. This inference is performed by combining minorize-maximization and self-concordant optimization techniques. The major advantage of the the proposed LRHP algorithm is that it is able to scale-up to datasets much larger than previous state-of-the-art methods, while maintaining performances very close to state-of-the-art competitors in terms of prediction and inference accuracy on synthetic as well as real datasets.

The chapter is organized as follows. In Section 1.2, we formally define multivariate Hawkes processes as well as the associated log-likelihood maximization problem. Through section 1.3, we propose our first algorithm MEMIP. In subsection 1.3.1, we decompose the log-likelihood on a basis of memoryless triggering kernels. In subsection 1.3.2, we develop two novel algorithms for exact as well as fast approximate maximization of the log-likelihood, analyze their complexity and show numerical convergence results based on the properties of self-concordant functions. In section 1.3.3, we show that MEMIP significantly improves over state of the art on both synthetic and real world data sets for the tasks of predicting future events as well as estimating underlying dynamics of the Hawkes process. Through Section 2, we propose the model LRHP that further reduces the complexity to $O(d\sum_{h=1}^{H} n_h)$. In subsection 1.2.1, we project the original dimensions in a low-rank space and decompose the triggering functions over a basis of exponential kernels. In subsection 1.2.2, we develop our new inference algorithm LRHP and show that its theoretical complexity is lower than state-of-the-art. In subsection 1.2.3, we empirically prove that LRHP outperforms significantly the state-of-the-art in terms of computational efficiency, while maintaining a very high level of precision for the task of recovering triggering kernels as well as predicting future events.

2.2 Estimation of Multivariate Hawkes Processes

2.2.1 The Multivariate Hawkes Process

We consider a multivariate Hawkes process, that is a d-dimensional counting process $N(t) = \{N^u(t) \,:\, u = 1,\dots,d\}$ for which the rate of occurrence of each component $N^u(t)$ is defined by:

$$\lambda_u(t) = \left(\mu_u(t) + \sum_{v \in [1\dots d]} \sum_{t_v < t} g_{vu}(t - t_v) \right)_+ , \quad \forall u = 1,\dots,d \tag{2.1}$$

where $\mu_u(t)$ is the natural rate of occurrence of events along dimension u. Note that the occurrence of a given event affects stochastic rates of occurrence of every dimension. With an empty history, events of type u will occur as if they were drawn from a non-homogeneous

Poisson process of rate $\mu_u(t)$. The kernel function evaluation $g_{uv}(t-t_v)$ quantifies the change in the rate of occurrence of event u at time t caused by the realization of event v at time t_v. Following the intuition, we can characterize three situations depending on the values taken by the kernel function at a given time lapse s:

- *Excitation* corresponds to the case where we have $g_{vu}(s) > 0$, i.e. an event of type v is more likely to occur if an event of type u has occurred at a time distance of s.

- *Independence* is observed when $g_{vu}(s) = 0$, meaning that the realization of an event of type u has no effect on the rate of occurrence of an event of type v at time distance s.

- *Inhibition* takes place when $g_{vu}(s) < 0$, i.e. an event of type v is less likely to occur if an event of type u occurred at time distance s.

Such processes can be seen as a generalization over the common definition of multivariate Hawkes process where the kernels g_{uv} are non-negative and the component-wise background rate μ_u is often taken constant.

2.2.2 Log-Likelihood of Multivariate Hawkes Processes

Input Observations. We define a *realization* h of a multivariate point process by the triplet $T_h^-, T_h^+, (t_i^h, u_i^h)_{i\in[1...n_h]}$, where T_h^- and T_h^+ are respectively the beginning and the end of the observation period, and (t_i^h, u_i^h), for $i \in [1...n_h]$, is the sequence of the n_h events occurring during this period. In the rest of the chapter, we will assume we are given n i.i.d realizations of a multivariate Hawkes process. Without loss of generality, we will assume $\min_h(T_h^-) = 0$ and take $T = \max_h(T_h^+)$.

Expression of the Log-Likelihood. We first set $\Lambda = \{\lambda_u \,:\, u = 1,\dots,d\}$. For a general multivariate point process, the log-likelihood of the whole dataset $\mathscr{H}$ is given by (*e.g.* [49]):

$$\mathscr{L}(\Lambda, \mathscr{H}) = \sum_{u=1}^{d} \sum_{h\in\mathscr{H}} \int_{T_h^-}^{T_h^+} \ln(\lambda_u(s)) dN_h^u(s) - \sum_{u=1}^{d} \sum_{h\in\mathscr{H}} \int_{T_h^-}^{T_h^+} \lambda_u(s) ds \tag{2.2}$$

where $\int f(s)) dN_h^u(s) = \sum_{i=1}^{n_h} f(t_i^h) 1\left\{u_i^h = u\right\}$. In the case of a linear Hawkes process (2.1), we introduce $\Lambda = (M, G)$ where $M = \{\mu_u \,:\, u = 1,\dots,d\}$ and $G = \{g_{u,v} \,:\, u, v = 1,\dots,d\}$

and the log-likelihood can be rewritten as:

$$\mathscr{L}(M,G,\mathscr{H}) = \sum_{h\in\mathscr{H}}\sum_{i=1}^{n_h} \ln\left(\mu_{u_i^h}(t_i^h) + \sum_{j\,:\,t_j^h<t_i^h} g_{u_j^h,u_i^h}(t_i^h - t_j^h)\right)$$
$$- \sum_{u=1}^{d}\sum_{h\in\mathscr{H}} \int_{T_h^-}^{T_h^+} \left(\mu_u(s) + \sum_{j=1}^{n_h} 1\left\{u_j^h = u\right\}\, g_{u,u_j}(s-t_j)\right)_+ ds \tag{2.3}$$

As shown in [50], the estimator obtained by maximizing the log-likelihood is under mild conditions consistent, convergent and asymptotically normal. In the rest of this section, we will focus on the algorithmic problem of efficiently computing this estimator. Depending on the parametrization of triggering kernels g_{uv}, this log-likelihood may or may not be concave. For instance, in the widely used setting where the background rates μ_u are constant and the kernels g_{uv} are non-negative and fixed up to the normalization factor ν_{uv}, the log-likelihood is concave and can be relatively easily maximized. However, even for the simple case of non-negative exponential kernels $g_{uv}(t) = \nu_{uv}\exp(-\alpha_j t)$ where $\nu_{uv} \geq 0$ the product term $\nu_{uv}\exp(-\alpha_v t)$ makes the log-likelihood not concave with respect to α_v. Therefore, global convergence of maximization methods is not guaranteed anymore.

2.3 Markovian Estimation of Mutually Interacting Processes (MEMIP)

2.3.1 Approximations of Multivariate Hawkes Processes on a Basis of Exponential Triggering Kernels

A K-approximation of the Multivariate Hawkes Process. For a given multivariate Hawkes process $\Lambda = (M,G)$, we consider finite approximations of the components of the rates of occurrence μ_u and g_{uv}. We first introduce the following functions:

$$\forall y \in [-\ln(T)/\alpha, 1], \quad \nu_u(y) = \mu_u(-\ln(y)/\alpha) \quad \text{and} \quad f_{uv}(y) = g_{uv}(-\ln(y)/\alpha)$$

and we use Bernstein-type polynomial approximations of order K for ν_u and f_{uv}: there exist coefficients $X_{uv,k}^K$ such that

$$\forall y \in [-\ln(T)/\alpha, 1], \quad \widehat{\nu}^K(y) = \sum_{k=0}^{K} X_{u0,k}^K y^k \quad \text{and} \quad \widehat{f}_{uv}^K(y) = \sum_{k=0}^{K} X_{uv,k}^K y^k \,.$$

These polynomial approximations are known to converge with a polynomial rate for smooth functions (with first r derivatives continuously differentiable) and geometric rate for analytic functions (see below). The K-approximation considered in this section relies on a simple change of variable in the Bernstein approximations by setting: $y = \exp(-\alpha t)$. We can now introduce the linear approximation of a multivariate Hawkes process with exponential kernels:

$$\forall t \in [0,T], \quad \widehat{\mu}^K(t) = \sum_{k=0}^{K} X^K_{u0,k} \exp(-k\alpha t) \quad \text{and} \quad \widehat{g}^K_{uv}(t) = \sum_{k=0}^{K} X^K_{uv,k} \exp(-k\alpha t) \ .$$

Classical arguments from approximation theory [51] lead to the following proposition.

Proposition 4. *For any function Ψ defined over $[0,T]$, we consider the supremum norm $||\Psi||_{T,\infty} = \sup_{t\in[0,T]} |\Psi(t)|$. The K-approximations $(\widehat{\mu}^K_u)_{K\geq 1}$ and $(\widehat{g}^K_{uv})_{K\geq 1}$ converge in supremum norm towards true functions μ_u and g_{uv} at the following rates:*

1. *if μ_u is C^r, $\left|\left|\mu_u(t) - \widehat{\mu}^K_u(t)\right|\right|^T_\infty = O(1/K^r)$*
2. *if μ_u is analytic, $\left|\left|\mu_u(t) - \widehat{\mu}^K_u(t)\right|\right|^T_\infty = O(\exp(-K))$*
3. *if g_{uv} is C^r, $\left|\left|g_{uv}(t) - \widehat{g}^K_{uv}(t)\right|\right|^T_\infty = O(1/K^r)$*
4. *if g_{uv} is analytic, $\left|\left|g_{uv}(t) - \widehat{g}^K_{uv}(t)\right|\right|^T_\infty = O(\exp(-K))$.*

Another property of the approximated multivariate Hawkes process is the Markov property of the counting process. We set $\widehat{N}^K(t)$ the d-dimensional Hawkes process uniquely defined by $\widehat{\lambda}^K = (\widehat{\mu}^K_u, \widehat{g}^K_{uv})_{u,v}$.

Proposition 5. *Assume that the empirical estimate $\widehat{N}^K(t)$ of the multivariate Hawkes process is obtained after i.i.d. realizations of $N(t)$ over the time interval $[0,T]$. There exists $(\widehat{\ell}^0, \widehat{\ell}^1, \ldots, \widehat{\ell}^K)$ such that:*

$$\forall u \in \{1,\ldots,d\} \ , \quad \widehat{\lambda}^K(t) = \sum_{k=0}^{K} \left(\widehat{\ell}^k(t)\right)_+$$

and $(\widehat{N}^K(t), \widehat{\ell}^0(t), \widehat{\ell}^1(t), \ldots, \widehat{\ell}^K(t))$ is a Markov Process on $\mathbb{N}^d \times \mathbb{R}^{d(K+1)}$.

The proof results from the following decomposition of each occurrence rate in the approximation: $\forall u \geq 1$,

$$\widehat{\lambda}_u^K(t) = \left(X_{u0,0}^K + \sum_{k=1}^{K} \left(X_{u0,k}^K \exp(-k\alpha t) + \sum_{v\,:\,t_v<t} X_{uv,(k-1)}^K \exp(-k\alpha(t-t_v)) \right) + \sum_{v\,:\,t_v<t} X_{uv,K}^K \exp(-(K+1)\alpha(t-t_v)) \right)_+$$

Markov property is then a direct consequence of the dynamics of the functions $\widehat{\ell}_u^k(t)$: they decay at rate $\exp(-k\alpha t)$ and jump by $X_{uv,(k-1)}^K$ whenever an event of type v occurs. As they entirely determine the stochastic rate which determines the conditional probability distribution of $\widehat{N}^K(t)$, the conditional probability distribution of future states of the process $(\widehat{N}^K(t), \widehat{\ell}^0(t), \widehat{\ell}^1(t), ...\widehat{\ell}^K(t))$ is uniquely determined by the present state.

A New Decomposition of the Log-Likelihood. The algorithms proposed in this section rely on a novel expression of the log-likelihood over a basis of triggering kernels. We use exponential excitation functions to account for non-linearity but our algorithms benefit from the properties of linear approximations. Based on the expression of the log-likelihood for general linear multivariate Hawkes process (2.3), we introduce the following notation to discover the specific expression for the K-approximation based on exponential triggering functions: $\forall u, v = 1, \ldots, d$, $\forall k = 1, \ldots, K$, $\forall h \in \mathscr{H}$, $\forall i = 1, \ldots, n_h$,

$$A_{uv,k}^{K,h,i} = \sum_{j\,:\,t_j^h<t_i^h} 1\left\{u_i^h = v, u_j^h = u\right\} \exp\left(-(k+1\{u>0\})\alpha(t_i^h - t_j^h)\right) \tag{2.4}$$

$$B_{0v,k}^{K,h}(s) = \exp(-k\alpha s) \tag{2.5}$$

$$B_{uv,k}^{K,h}(s) = \sum_{j\,:\,t_j^h<s} 1\left\{u_j^h = v\right\} \exp(-(k+1)\alpha(s - t_j^h)) \tag{2.6}$$

The key expression of the approximate log-likelihood can then be derived by plugging-in the previous notations and replacing the intrinsic parameters (M, G) by the linear coefficients X^K:

$$\mathscr{L}^K(X^K, \mathscr{H}) = \sum_{h\in\mathscr{H}} \sum_{i=1}^{n_h} \ln(A^{K,h,i} X^K) - \sum_{h\in\mathscr{H}} \int_0^{T_h} \left(\sum_{i=1}^{n_h} B^{K,h}(s) X^K \right)_+ ds \tag{2.7}$$

Note that the dependency of $\mathscr{L}^K$ on the history $\mathscr{H}$ is entirely expressed by vectors $(A^{K,h,i})_{h\in\mathscr{H}, i\in[1\ldots n_h]}$ and $(B^{K,h}(s))_{h\in\mathscr{H}, s\in[0,T]}$. An important feature of the approximate log-

likelihood expressed in the parameter space defined by linear decomposition onto bases of exponential triggering kernels is given in the following proposition.

Proposition 6. *The function* $X \to \mathscr{L}^K(X, \mathscr{H})$ *is concave.*

From there, we have a complete roadmap for the design of algorithms estimating the parameters of multidimensional Hawkes processes: the last proposition indicates that a proxy of the log-likelihood (2.3) can be globally maximized with tools of convex analysis. Moreover, thanks to the approximation rates of convergence (Proposition 4), triggering kernels can be accurately estimated for large K through maximization of the new objective (2.7). Finally, the Markov property is an important feature that will allow us to construct the vectors $(A^{K,h,i})$ and $(B^{K,h})$ with linear complexity.

2.3.2 Markovian Algorithms for the Estimation of Triggering Kernels

Computational tractability of algorithms on large data sets depends on the algorithmic complexity in the dominating dimensions of the problem. For realizations of multivariate Hawkes processes, dominating dimensions are almost always the total number of events $N = \sum_{h \in \mathscr{H}} n_h$ and the time of observation T. Indeed, it would be unrealistic to try to learn d^2 non-parametric functions in an infinite dimensional space with only N observations without the condition $N \gg d^2$. In the rest of this section, we will therefore focus on constructing two algorithms with no more than linear complexity in N and T.

Exact Maximization of the Approximated Log-Likelihood. Vectors $(A^{K,h,i})_{h \in \mathscr{H}, i \in [1 \dots n_h]}$ and $(B^{K,h}(s))_{h \in \mathscr{H}, s \in [0,T]}$ can be constructed in a single pass through the data thanks to **Algorithm 1**.

Complexity of Algorithm 1. With $M = T/dt$ the number of discretizations steps, construction of vectors $(A^{K,h,i})$ and $(B^{K,h}(s))$ has thus a complexity of $O(N+M)$. As each log-likelihood evaluation (2.7) requires $2N+M$ scalar products computations, various optimization techniques can be used to find the global maximum of $X \to \mathscr{L}^K(X, \mathscr{H})$ in $O(N+M)$ operations. On the contrary, a nonmarkovian estimator, even linear, would need at each time t to compute the values of triggering kernels between current time and all preceding occurrence times, thus leading to a $O(\sum_h n_h^2)$ complexity. This construction is thus very often the bottleneck of the whole maximization procedure.

Relaxed Version of the Log-Likelihood. While the previous paragraph exposes a fully tractable method to estimate the triggering kernels for potentially large data sets, we now

Algorithm 1 Algorithm for construction of vectors $(A^{K,h,i})$ and $(B^{K,h}(s))$

Initialize $i = 0$ and fix a time step dt
for all h **do**
  Initialize $(C^k_{uv} = 0)_{u\geq 1, v\geq 1}$; $t = T_h^-$; $(D^k_{uv}(T_h^-) = 1_{\{u=0\}})_{u\geq 0, v\geq 1}$
  while $t < T_h^+$ **do**
    $t \leftarrow t + \delta t = \min(t + dt, t_i)$
    for all k,u,v **do**
      $C^k_{uv} \leftarrow C^k_{uv} \exp(-(k + 1\{u > 0\}\alpha\delta t)$, $D^k_{uv} \leftarrow D^k_{uv} \exp(-(k + 1\{u > 0\}\alpha\delta t)$
      $B^{K,h}_{uv,k}(t) \leftarrow D^k_{uv}$
    end for
    if $t = t_i$ **then**
      for all k,u **do**
        $A^{K,h,i}_{uv,k} \leftarrow C^k_{uu_i}$
      end for
      for all k,v **do**
        $C^k_{u_i v} \leftarrow C^k_{u_i v} + 1$, $D^k_{u_i v} \leftarrow D^k_{u_i v} + 1$
      end for
      $i \leftarrow i + 1$
    end if
  end while
end for

develop an approximate algorithm called MEMIP, for Markovian Estimation of Mutually Interacting Processes, that leads to a substantial speed-up, as well as theoretical guarantees in terms of efficiency. For this purpose, we approximate the log-likelihood $\mathscr{L}^K(M, G, \mathscr{H})$ by dropping the positive part in log-likelihood (2.3), i.e.

$$\widetilde{\mathscr{L}^K}(M, G, \mathscr{H}) = \sum_{h\in\mathscr{H}} \left(\sum_{i=1}^{n_h} \ln\left(\mu_{u_i^h}(t_i^h) + \sum_{j\,:\,t_j^h < t_i^h} g_{u_j^h, u_i^h}(t_i^h - t_j^h) \right) \right.$$
$$\left. - \sum_{u=1}^{d} \int_{T_h^-}^{T_h^+} \left(\mu_u(s) + \sum_{j=1}^{n_h} 1\left\{u_j^h = u\right\} g_{u,u_j}(s - t_j) \right) ds \right) \tag{2.8}$$

which can be rewritten:

$$\widehat{\mathscr{L}^K}(X^K, \mathscr{H}) = \sum_{h\in\mathscr{H}} \left(\sum_{i=1}^{n_h} \ln(A^{K,h,i} X^K) \right) - \widehat{B}^K X^K \tag{2.9}$$

where $\widehat{B}^K_{uv,k} = \sum_{h\in\mathscr{H}} \sum_{j=1}^{n_h} 1\left\{u_j^h = v\right\} \int_{T_h^-}^{T_h^+} \exp(-k\alpha(s - t_j^h))$.

Although $\widehat{\mathscr{L}^K}(X,\mathscr{H})$ is an upper bound of the actual log-likelihood and it is not clear at first sight why its maximization should lead to large values of $\mathscr{L}^K(X,\mathscr{H})$, we point out that the difference $\widehat{\mathscr{L}^K}(X,\mathscr{H}) - \mathscr{L}^K(X,\mathscr{H})$ is only caused by intervals where there exists $u \in [1...d]$ such that $\widehat{\lambda}_u^K(t) = 0$. But maximizers of $\widehat{\mathscr{L}^K}(X,\mathscr{H})$ are very unlikely to exhibit wide range of negative values in their triggering kernels because any single event realization with a predicted nonpositive stochastic rate yields $\widehat{\mathscr{L}^K}(X,\mathscr{H}) = -\infty$. Therefore, we assume we can rely on this approximation in order to construct fast algorithms.

2.3.3 MEMIP: a Learning Algorithm for Fast Log-Likelihood Estimation

Since the gradient and the hessian matrix of $X \mapsto \widehat{\mathscr{L}^K}(X,\mathscr{H})$ can be computed analytically and their size does not depend on N, we derive the proposed algorithm MEMIP on the base of successive damped Newton optimization steps, i.e. using Newton method with backtracking linesearch (see e.g. [52]). In the sequel, we denote by NewtonArgMax(f, x_0) the result of such a maximization of function f using with starting point x_0. using The main idea is to construct recursively a sequence $(\widehat{X^1}...\widehat{X^K})$ of maximizers of functions $(\widehat{\mathscr{L}^k})_{k\in[1...K]}$ by using NewtonArgMax$(\widehat{\mathscr{L}^{k-1}}, \widehat{W}^{k-1})$ as the starting point $\widehat{W}^k$ of maximization of $\widehat{\mathscr{L}^k}$. From the estimated sequence $(\widehat{X}^1...\widehat{X}^K)$, the best value of k can be estimated by cross-

Algorithm 2 Algorithm (MEMIP) for learning background rates and triggering kernels of a multivariate Hawkes process

input Mapping parameter $\alpha > 0$, maximal polynomial degree K, starting point $\widehat{W}^1 \in \mathbb{R}^{d(d+1)}$
  Construct $(A^{K,h,i})$ and B^K according to $O(N)$ modified version of **Algorithm 1**
  $\widehat{X}^1 \leftarrow$ NewtonArgMax$(\widehat{\mathscr{L}^1}, \widehat{W}^1)$
  for $k \in [2...K]$ **do**
    $\widehat{W}^k = 0$
    for $j \in [1...k-1], u \in [1...d], v \in [0...d]$ **do**
      $\widehat{W}^k_{uv,j} = \widehat{X}^{k-1}_{uv,j}$
    end for
    $\widehat{X}^k \leftarrow$ NewtonArgMax$(\widehat{\mathscr{L}^k}, \widehat{W}^k)$
  end for

validation or various other model selection techniques. Interestingly, $A^{k,h,i} = (A^{K,h,i}_{\bullet,j})_{j\in[1...k]}$ and $B^k = (B^K_{\bullet,j})_{j\in[1...k]}$, and therefore only $(A^{K,h,i})_{h\in\mathscr{H},i\in[1...n_h]}$ and B^K need to be computed.

Complexity of Algorithm 2. We obtain two substantial computational speed-ups compared to exact log-likelihood maximization. First, time discretization is no longer needed for the construction of B^K. Thus, vectors $(A^{K,h,i})$ and B^K can be constructed with the same procedure

than **Algorithm 1** except that updates are made only on time occurrence of events. Therefore, construction complexity is $O(N)$. Similarly, approximate log-likelihood evaluations are also of complexity $O(N)$. Secondly, the approximate log-likelihood is separable by type of event u such that $\widehat{\mathscr{L}^K} = \sum_{u=1}^{d} \widehat{\mathscr{L}_u^K}$ where $\widehat{\mathscr{L}_u^K}$ only depends on the background rate μ_u and triggering kernels $(g_{uv})_{v\in[1...d]}$. Maximization can thus be parallelized across the different dimensions. Note that because of the Hessian inversion at each Newton step, complexity in d of maximization of $\widehat{\mathscr{L}_u^K}$ is $O(d^3)$ for any u, which yields a $O(d^4)$ overall complexity. In cases where $N \gg d^2$ but $d^4 > N$, the use of quasi-Newton method might therefore be preferable. For instance, BFGS method enjoys superlinear convergence [53], and would lead to a $O(d^3)$ overall complexity since the maximization of each $\widehat{\mathscr{L}_u^K}$ requires $O(d^2)$ operations.

Self-Concordance Property and Numerical Convergence of MEMIP. Problem (2.9) can be solved by various optimization techniques, such as L-BFGS-B [54] or stochastic gradient descent. **Algorithm 2** is actually based on the concept of *self-concordance* [55] that we apply to function $X \mapsto -\widehat{\mathscr{L}^k}(X, \mathscr{H})$. Self-concordant functions are, along with strongly-convex functions with Lipschitz-continuous Hessian matrices, a very important class of functions for which non-asymptotic upper bounds of the number of Newton steps necessary to reach precision ε is known. More specifically, the following property holds:

Proposition 7. *The d(d+1)-dimensional MEMIP estimates $\widehat{X}^k$ for $k \in \{1,\dots,K\}$ with initialization vector $\widehat{W}^1$ satisfy the following condition: there exists $C > 0$ such that for any $\varepsilon > 0$ and $k \in [1...K]$, we have $|\widehat{\mathscr{L}^k}(\widehat{X}^k, \mathscr{H}) - \sup_X(\widehat{\mathscr{L}_k}(X, \mathscr{H}))| \leq \varepsilon$ after at most $C\big(\sup_X(\widehat{\mathscr{L}_K}(X, \mathscr{H})) - \widehat{\mathscr{L}_1}(\widehat{W}^1, \mathscr{H})\big) + K(\log_2\log_2(1/\varepsilon) + C\varepsilon)$ Newton iterations.*

Proof of Proposition 4. This proof heavily relies on the following lemma that bounds the number of damped Newton steps necessary to minimize a self-concordant function up to a precision ε.

Lemma 1. *The series of estimates $x_n \in \mathbf{R}^d$ for $n \geq 0$ obtained by successive Newton iterations with backtracking line search parameters α and β and initialization vector x_0 satisfy the following condition: there exists $C(\alpha,\beta) > 0$ such that the total number of Newton iterations needed to minimize a self-concordant function f up to a precision ε is upper bounded by $C(\sup(f) - f(x_0)) + \log_2\log_2(\frac{1}{\varepsilon})$.*

Self-concordance of functions $(-\widehat{\mathscr{L}_k})_{k\in[1...K]}$ is a direct consequence of self-concordance on $\mathbf{R}_+^*$ of $f : x \mapsto -\ln(x)$ and affine invariance properties of self-concordant functions. By applying the aforementioned lemma to function $-\widehat{\mathscr{L}_k}$ and starting point $\widehat{W}^k$ at each Newton

optimization, we get the bound

$$C\sum_k \left(\sup_X(\widehat{\mathscr{L}_k}(X,\mathscr{H})) - \widehat{\mathscr{L}^k}(\widehat{W}^k,\mathscr{H})\right) + K\log_2\log_2(1/\varepsilon) \tag{2.10}$$

But we also have $\widehat{\mathscr{L}^k}(\widehat{W}^k,\mathscr{H}) = \widehat{\mathscr{L}^{(k-1)}}(\widehat{W}^k,\mathscr{H}) = \widehat{\mathscr{L}^{(k-1)}}(\widehat{X}^{k-1},\mathscr{H})$ where the first equality holds because for any u, v, $\widehat{W}^k_{uv,k} = 0$ and the second because for any u, v, $j \leq k-1$, $\widehat{W}^{k-1}_{uv,j} = \widehat{X}^{k-1}_{uv,j}$. But for any $k \geq 2$, $\widehat{\mathscr{L}^{k-1}}(\widehat{X}^{k-1},\mathscr{H}) \geq \sup_X(\widehat{\mathscr{L}_{k-1}}(X,\mathscr{H})) - \varepsilon$. Therefore, using the notation $\sup_X(\widehat{\mathscr{L}_0}(X,\mathscr{H})) = \widehat{\mathscr{L}_1}(\widehat{W}^1,\mathscr{H})$, the bound reformulates as

$$C\sum_{k=1}^{K}\left(\sup_X(\widehat{\mathscr{L}_k}(X,\mathscr{H})) - \sup_X(\widehat{\mathscr{L}_{k-1}}(X,\mathscr{H}))\right) + K(\log_2\log_2(1/\varepsilon) + C\varepsilon) \tag{2.11}$$

which proves Proposition 4. □

Remark. The previous proposition emphasizes the key role played by the starting point $\widehat{W}^1$ in the rate of convergence of Newton-like methods. In our case, a good choice is for instance to select it by classical non-negative maximization techniques for objectives of type (2.9) (see *e.g* [56]). Because these methods are quite fast, they can also be used for steps $k \in [2...K]$ in order to provide an alternative starting point $\widehat{W}^k_+$. The update $\widehat{X}^k$ is then given by either NewtonArgMax$(\widehat{\mathscr{L}^k},\widehat{W}^k)$ or NewtonArgMax$(\widehat{\mathscr{L}^k},\widehat{W}^k_+)$ depending on which proposed update yields the highest log-likelihood.

2.3.4 Experimental Results on simulated and MemeTracker datasets

We first evaluate MEMIP on realistic synthetic data sets. We compare it to

- fixed exponential kernels, i.e. for a given $\alpha > 0$ and any $u,v \in [1...d]$, $g_{uv}(t) = \nu_{uv}\exp(-\alpha t)$.
- MMEL [7], a state-of-the-art algorithm relying on a low-rank decomposition of each triggering kernels g_{uv} over a dictionary of basis functions $(h_c)_{c\in[1...C]}$ in the following manner: $g_{uv}(t) = \nu^c_{uv}g_c(t)$. The principle of MMEL is to alternate between estimation of the decomposition parameters ν^c_{uv} and non-parametric estimation of the basis functions g_c, in a EM-like fashion.

We will show in the next section that MEMIP performs significantly better in terms of prediction and triggering kernels recovery than both baselines.

Data Generation. We simulate multivariate Hawkes processes by *Ogata modified thinning algorithm* (see e.g. [36]). Since each occurrence can potentially increase stochastic rates of

all events, special attention has to be paid to avoid *explosion*, i.e. the occurrence of an infinite number of events on a finite time window. In order to avoid such behavior, our simulated data sets verify the sufficient non-explosion condition $\rho(\Gamma) < 1$ where $\rho(\Gamma)$ denotes the spectral radius of the matrix $\Gamma = (\int_0^\infty |g_{uv}(t)dt|)_{uv}$ (see *e.g* [49]). We perform experiments on three different simulated data sets where triggering kernels are taken as

$$g_{uv}(t) = \nu_{uv} \frac{\sin\left(\frac{2\pi t}{\omega_{uv}} + \frac{\pi}{2}((u+v) \bmod 2)\right) + 2}{3(t+1)^2} \tag{2.12}$$

We sample the periods ω_{uv} from a uniform distribution over $[1,10]$. Absolute values of normalization factors ν_{uv} are sampled uniformally from $[0,1/d[$ and their sign is sampled from a Bernoulli law of parameter p. Except for the toy data set, background rates μ_v are taken constant and sampled in $[0,0.001]$. An important feature of this choice of triggering kernels and parameters is that resulting Hawkes processes respect the aforementioned sufficient non-explosion condition. For quantitative evaluation, we simulate two quite large data sets (1) $d = 300$, $p = 1$ (2) $d = 300$, $p = 0.9$. Thus, data set (1) contains realizations of purely mutually-exciting processes whereas data set (2) has 10% of inhibitive kernels. For each data set, we sample 10 sets of parameters $(\omega_{uv}, \nu_{uv})_{u\geq 1, v\geq 1}$,$(\mu_v)_{v\geq 1}$ and simulate 400,000 i.i.d realizations of the resulting Hawkes process over $[0,20]$. The first 200,000 are taken as training set and the remaining 200,000 as test set.

Evaluation Metrics. We evaluate the different algorithms by two metrics: (a) *Diff* a normalized L^2 distance between the true and estimated triggering kernels, defined by

$$\text{Diff} = \frac{1}{d^2} \sum_{u=1}^{d} \sum_{v=1}^{d} \frac{\int (\widehat{g}_{uv} - g_{uv})^2}{\int \widehat{g}_{uv}^2 + \int g_{uv}^2}, \tag{2.13}$$

(b) *Pred* a prediction score on the test data set defined as follows. For each dimension $u \in [1...d]$ and occurrence i in the test set, probability for that occurrence to be of type u is given by $P_i^{true}(u) = \frac{\lambda_u(t_i)}{\sum_{v=1}^d \lambda_v(t_i)}$. Thus, defining $AUC(d,P)$ the area under ROC curve for binary task of predicting $(1_{\{u_i=u\}})_i$ with scores $(P_i^{true}(d))_i$ and $(P_i^{model}(d))_i$ the probabilities estimated by the evaluated model, we set

$$\text{Pred} = \frac{\sum_{u=1}^d (AUC(d, P^{model}) - 0.5)}{\sum_{u=1}^d (AUC(d, P^{true}) - 0.5)} \tag{2.14}$$

Baselines. We compare MEMIP to (a) **MMEL** for which we try various sets of number of base kernels, total number of iterations and smoothing hyperparameter, (b) **Exp** the widely

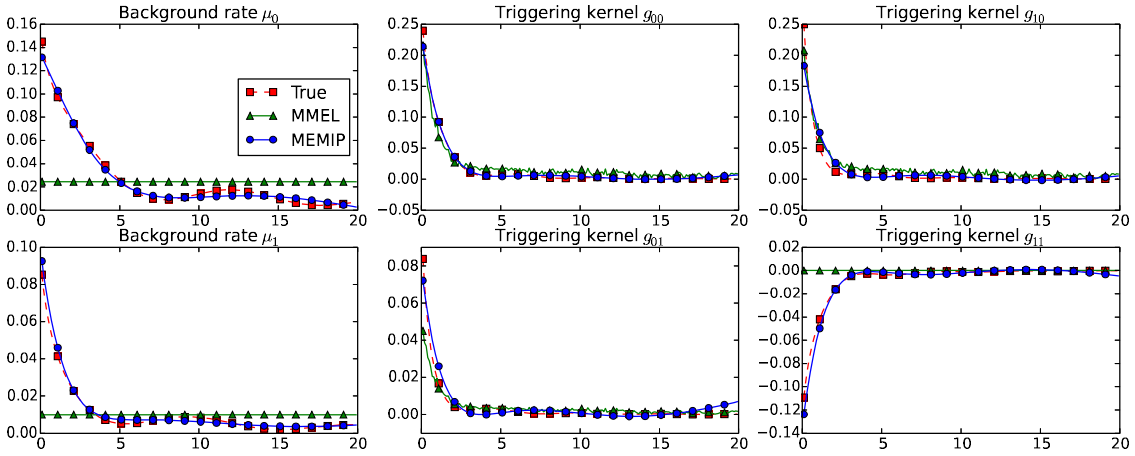


Fig. 2.1 Triggering kernels and background rates w.r.t. time (abscissa) for toy data set estimated by MEMIP and MMEL algorithms vs true triggering kernels and background rate

used setting where $g_{uv}(t) = \nu_{uv} \exp(-\alpha t)$ and only ν_{uv} are estimated from the data. In order to give this baseline more flexibility and prediction power, we allow negative values of ν_{uv}. We train three different versions with $\alpha \in \{0.1, 1.0, 10.0\}$.

Results - Part 1: Visualization on a Toy Dataset. In order to demonstrate the ability of MEMIP to discover the underlying dynamics of Hawkes processes even in presence of inhibition and varying background rates, we construct the following bidimensional toy data set. Amongst the four triggering kernels, g_{11} is taken negative and background rates are defined by $\mu_0 = \frac{cos(\frac{2\pi t}{\omega_0})+2}{1+t}$ and $\mu_1 = \frac{sin(\frac{2\pi t}{\omega_1})+2}{1+t}$ with parameters ω_0 and ω_1 sampled in $[5, 15]$. We sample a set of parameters $(\omega_{uv}, \nu_{uv})_{u\geq 1, v\geq 1}$,$(\mu_v)_{v\geq 1}$ and simulate 200,000 i.i.d realizations of the resulting Hawkes process. From Fig. **2.1**, we observe that both compared methods MEMIP and MMEL accurately recover non-negative triggering kernels g_{00}, g_{01} and g_{10}. However, MEMIP is also able to estimate the inhibitive g_{11} whereas MMEL predicts $g_{11} = 0$. Varying background rates μ_0 and μ_1 are also well estimated by MEMIP, whereas by construction MMEL and Exp only return constant values $\bar{\mu}_0$ and $\bar{\mu}_1$.

Results - Part 2: Prediction Score. In order to evaluate **Pred** score of the competing methods on the generated data sets, we remove for each model the best and worst performance over the ten simulated processes, and average **Pred** over the eight remaining ones. Empirical 10% confidence intervals are also indicated to assess statistical significance of the experimental results. From Table **1**, we observe that MEMIP significantly outperforms the competing baselines for both data sets. Prediction rates are quite low for all methods which indicates a rather difficult prediction problem, as $90,000$ non-parametric functions are indeed to be estimated from the data. In Fig. **2.2** , we study the sensitivity of **Pred** score to

Table 2.1 Pred score for prediction of the type of next event on simulated data sets

Dataset	MEMIP	MMEL	Exp
(1) d=300,p=1	$0.288 \in [0.258, 0.310]$	$0.261 \in [0.250, 0.281]$	$0.255 \in [0.236; 0.278]$
(2) d=300,p=0.9	$0.287 \in [0.266, 0.312]$	$0.261 \in [0.241, 0.280]$	$0.256 \in [0.242, 0.280]$

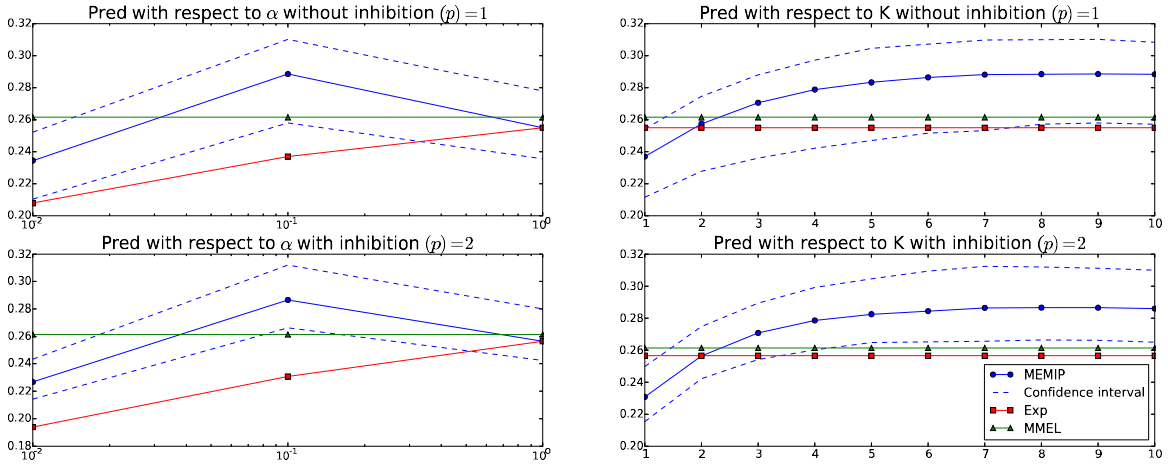


Fig. 2.2 Sensitivity to hyperparameters α (left) and K(right) for **Pred** score of MEMIP algorithm, compared to Exp and MMEL baselines on non-inhibitive simulated data set (above) and simulated data set with 10 % inhibitive kernels (below)

α and K for simulated data sets (1)(above) and (2)(below). Left plots show MEMIP and Exp **Pred** score with respect to α, as well as best MMEL average score across a broad range of hyperparameters. Empirical 10% confidence intervals are also plotted in dashed line. We see that MEMIP gives good results in a wide range of values of α, and outperforms the exponential baseline for all values of α. Right plots show MEMIP **Pred** score with respect to K for $\alpha = 0.1$, as well as best Exp and MMEL average score. We see that MEMIP achieves good prediction results for low values of K, and that taking $K > 10$ is not necessary. For very large values of α, we also note that MEMIP and Exp baseline are the same, because the optimal choice of K for MEMIP is $K = 1$.

Results - Part 3: Accuracy of Kernel Estimation. Besides having a greater prediction power, we observe in Table **2** that MEMIP is also able to estimate the true values of triggering kernels more accurately on both data sets. In Fig. **2.3**, we study the sensitivity of **Diff** score to α and K for data sets (1)(above) and (2)(below). We see that the variance of **Diff** score is very low for MEMIP, and its fitting error is significantly lower than those of the baselines at level 10%.

Table 2.2 Diff score for triggering kernels recovery on simulated data sets

Dataset	MEMIP	MMEL	Exp
(1) d=300,p=1	$0.759 \in [0.755, 0.768]$	$0.807 \in [0.803, 0.814]$	$0.791 \in [0.788, 0.800]$
(2) d=300,p=0.9	$0.803 \in [0.793, 0.810]$	$0.839 \in [0.833, 0.844]$	$0.830 \in [0.818, 0.836]$

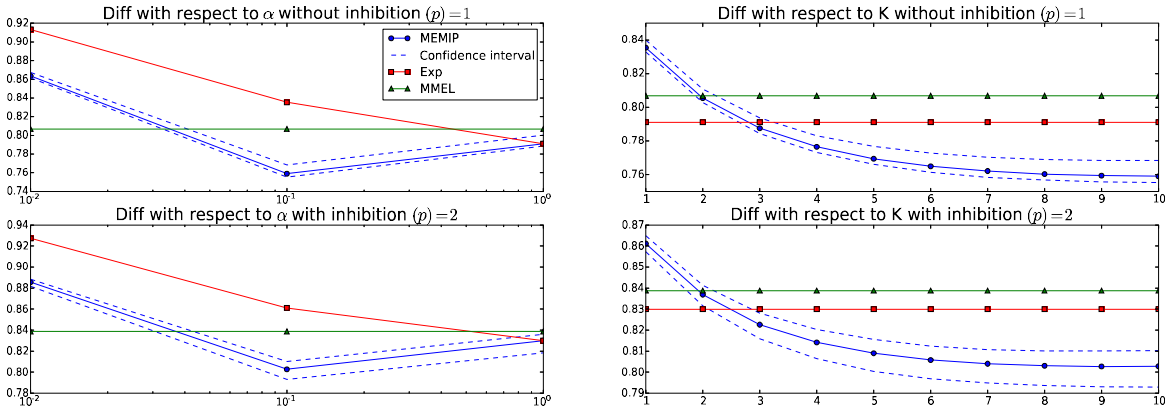


Fig. 2.3 Sensitivity to hyperparameters α (left) and K(right) for **Diff** score of MEMIP algorithm, compared to Exp and MMEL baselines on non-inhibitive simulated data set (above) and simulated data set with 10 % inhibitive kernels (below)

Comparison to Related Work. The closest work to ours is the algorithm MMEL derived in [7] by Zhou, Zha and Song. MMEL decomposes the triggering kernels on a low-rank set of basis functions, and makes use of EM-like methods in order to maximize the log-likelihood. Compared to MMEL, the proposed algorithm MEMIP enjoy three main improvements: 1) $O(N)$ complexity instead of $O(N^2)$, 2) global convergence of log-likelihood maximization, 3) the ability to learn negative projection coefficients $X_{uv,k}$ as well as varying background rates. Experimental results also suggest that MEMIP may outperform MMEL significantly even for non-inhibitive data set. Actually, even in purely mutually-exciting settings, these two algorithms can exhibit quite different behaviors due to their smoothing strategies. Indeed, because the log-likelihood (2.1) can be made arbitrarily high by the sequence of functions $(g^n_{uv})_{n\in N}$ defined by $g^n_{uv}(t) = n1_{\{t\in T_{uv}\}}$ where $T_{uv} = \{t_v - t_u \mid (t_u < t_v \wedge (\exists h \in \mathscr{H} \mid (t_v, v) \in h \wedge (t_u, u) \in h))\}$, smoothing is mandatory when learning triggering kernels by means of log-likelihood maximization. Using a L^2 roughness norm penalization $\alpha \int_0^T g'^2$, MMEL can face difficult dilemmas when fitting power-laws fastly decaying around 0 : either under-estimating the rate when it is at its peak or lowering the smoothness parameter and being vulnerable to overfitting. On the contrary, MEMIP would face difficulties to perfectly fit periodic functions with a very small period, as the derivative of its order K estimates can only vanish $K-1$ times.

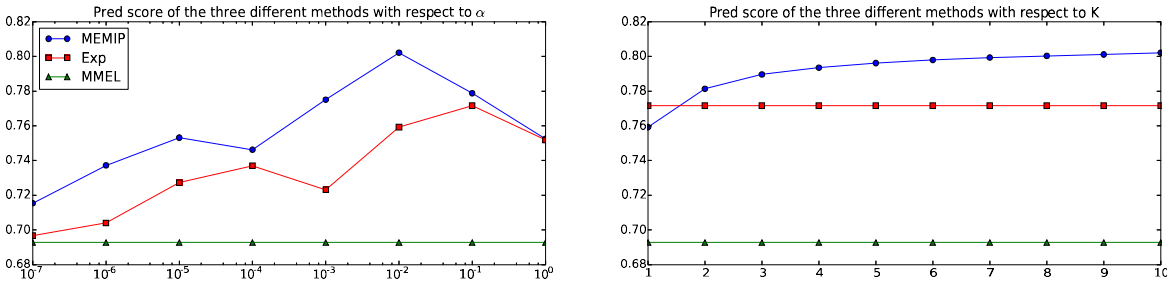


Fig. 2.4 Sensitivity to hyperparameters α (left) and K(right) for prediction score of MEMIP algorithm, compared to Exp and MMEL baselines on MemeTracker data set

Experiment on the MemeTracker Data Set. In order to show that the ability to estimate inhibitive triggering kernels and varying background rates yields better accuracy on real-world data sets, we compare the proposed method MEMIP to different baselines on the MemeTracker data set, following the experience plan exposed in [7]. MemeTracker contains links creation between some of the most popular websites between August 2008 and April 2009. We extract link creations between the top 100 popular websites and define the occurrence of an event for the i^{th} website as a link creation on this website to one the 99 other websites. We then use half of the data set as training data and the other half at test data on which each baseline is evaluated by average area under ROC curve for predicting future events. From Fig. **2.4**, we observe that the proposed method MEMIP achieves a better prediction score than both baselines. Left plot shows MEMIP and Exp prediction score with respect to α, as well as best MMEL score across a broad range of hyperparameters. We see that MEMIP gives good results in a very broad range of values of α, and significantly outperforms the exponential baseline for all values of α. Right plot shows MEMIP prediction score with respect to K for $\alpha = 0.01$, as well as best Exp and MMEL score. For $K = 10$, MEMIP achieves a prediction score of 0.8021 whereas best MMEL and Exp score are respectively 0.6928 and 0.7716. We note that, even for K as low as 3, MEMIP performs the prediction task quite accurately.

Posterior work in the literature. Since the publication of [33], the paper on which rely this section , it was shown in [57] that algorithms heavily inspired from MEMIP also outperform MMEL on more structured datasets where the matrix of interactions between the different dimensions of the Hawkes process is block-diagonal. Two interesting choices were made in this work: first, the objective defined in Eq. 2.9 was maximized using stochastic gradient descent, and specifically SVRG [58] which outperform the "vanilla" SGD algorithm in the presented experiments. Secondly, the author propose the following parametrization of

the triggering kernels:

$$\forall t \in [0,T], \quad \widehat{\mu}^K(t) = \sum_{k=0}^{K} X^K_{u0,k} \exp(-l^k \alpha t) \quad \text{and} \quad \widehat{g}^K_{uv}(t) = \sum_{k=0}^{K} X^K_{uv,k} \exp(-l^k \alpha t) \ ,$$

where $l > 1$ is an additional hyperparameter. The rationale behind these kernels is that they allow to better capture interactions taking place at different time scales. Indeed, if the triggering kernels are heavy-tailed (for instance non-negligible at the second scale as well as the millisecond scale), the number of basis functions in $\exp(-k\alpha t)$ needed to encode this behavior would be very large. However, the price to pay for this very promising idea is that theoretical convergence of the sequence of estimates towards the true triggering kernels is no longer guaranteed. Indeed, Müntz-Sasz theorem [59] categorizes the sequences ϕ_i such that x^{ϕ_i} spans a dense subset of $C([0,1])$ in the following way (see e.g [60]):

Proposition 8. *Let $(\phi_i)_{i\geq 1}$ a sequence of distinct numbers, such that $1 \leq \phi_i \to \infty$. Then, the set of functions $\{1, x^{\phi_1}, x^{\phi_2}, ...\}$ is fundamental in $C([0,1])$ if and only if $\sum_{i\geq 1} 1/\phi_i = \infty$.*

The choice of the parametrization should therefore be made depending on whatever is the most important between the precision of the estimate at one given time scale, or the ability to be able to take into account different time scales.

2.3.5 Operational application for real-time bidding

Industrial problem of interest. One of the main industrial goals of this PhD was to develop a method for predicting the internet users most likely to be interested in performing an action of interest , or *conversion*(*e.g* purchase, form completion...) in the next few days (attribution windows range typically from 1 to 30 days), in order to better allocate the advertising budget. The data at disposition consists in 4.1×10^8 users 1000mercis has recorded during the last three months (i.e. the internet user has visited a tagged website or has been printed a banner). The resulting dataset consists of 9.0×10^9 visits of URLs, with a total of $d = 9.9 \times 10^8$ visited URLs. The purpose of the scoring in to be able to invest more on people being more likely to convert.

Internet users modelization. We model the navigation histories of the different internet users as i.i.d. realizations of a common multivariate Hawkes process. The dimensions of the processes are composed of the conversion to predict (taken as the dimension 0) and visits on different clusters of URL (dimensions $1...r$). Once the inference has been performed by MEMIP algorithm, the probability of each user to convert between T and $T+t$ can be

computed by means of Monte-Carlo simulations. An interesting alternative is to simply use as a score the conditional rate of occurrence $\lambda_0(t)$, which is indeed proportional to the probability of conversion of each internet user in a small interval after T. This choice, made in 1000mercis implementation, has the major advantage to greatly reduce the complexity of the algorithm: indeed, we only have to infer μ_0 and the r triggering kernels g_{u0} for $u \in [1...r]$ and not the $r(r+1)$ functions encoding the interactions of the whole Hawkes process. However, it is not clear at first sight whether or not this short-term prediction is a good approximation for what will happen in several days. We will show in the experimental section that this solution works quite well in practice.

Preprocessing of URLs. Despite the linear complexity of this scoring method in the total number of URL visits and in the total number of visited URLs, the MEMIP algorithm is not directly applicable to the problem, due to the very large number of URLs. The most natural workaround is to find a way to cluster these URLs into r mutuallly exclusive classes with $r \ll d$. Ideally, these classes should be large enough so that a large proportion of the URLs belong to one of them, and sufficiently correlated to the conversion so that we select the user actions with the most predicting power. To do so, the main features to be considered are :

- URL structure / semantics
- navigation histories of users having visited the associated web page
- associated web page semantics (necessitates crawling of the URL)

In Sec. 2.4, we will see how the clustering can be performed simultaneously inside the Hawkes inference procedure, using only the internet users navigation histories. However, the operational choice made at 1000mercis was to focus on the URL structures in order to perform this segmentation. More specifically, we distribute the different URLs $u = 1...U$ on a tree where each node is a set of URLs represented by a descriptor of the form $domain/category_1/category_2/.../category_n/\%$ and contains all the URLs starting by the string $domain/category_1/category_2/.../category_n/$ (see Fig. 2.5).

Note that the set $\mathcal{V}_1$ of nodes at depth 1 is a partition of the set containing all the URLs, and more generally the set $\mathcal{V}_n$ of nodes at depth n is a partition of the set containing all the URLs of depth at least n. Arguably, a maximal depth td of the tree should be fixed for computational reasons. Then, for each node v, we consider the action "visit of a URL $u \in v$" as a candidate for a dimension of the multivariate Hawkes process. We describe each node by : 1) the total number of visits $vis(v)$ on the URLs belonging to the node v during the last three months, and 2) the "conversion rate" $\tau(v)$ of users having visited at

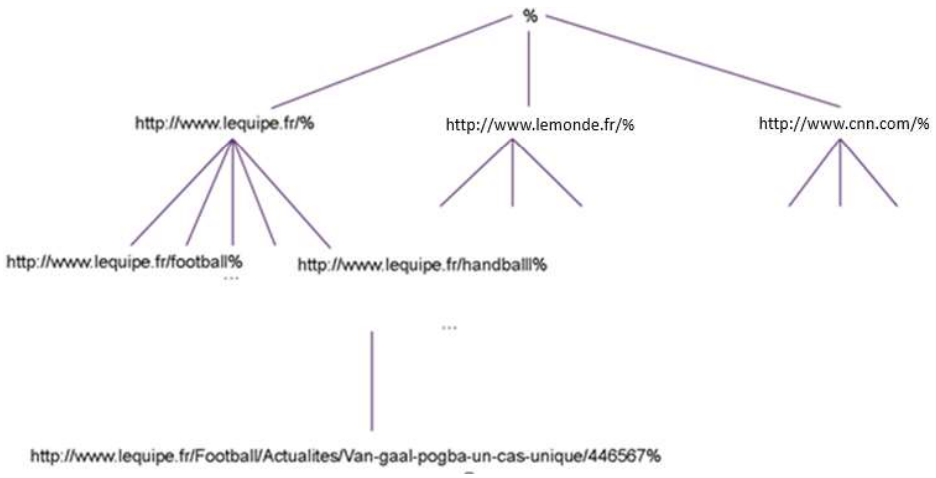


Fig. 2.5 Example of descriptor tree

least one of the URLs belonging to the node (i.e. number of users having converted after a visit divided by the number of users having visited). The selection procedure, summarized in Alg. 3, allows to select $r \in [r_{min}, r_{max}]$ nodes with minimum number of visits vis_{-} by choosing the right minimum conversion rate τ_{-}. Given these two parameters, the algorithm starts from the leafs (URL paths of length td) and see whether they meet the criterion on the minimum number of visits and conversion rates. If yes, it selects the node as dimension of the Hawkes processes and aliment a URL blacklist $\mathscr{U}$ with the corresponding URLs. Then, it evaluates the parent nodes only on the basis of the incremental URLs (i.e. if the node $domain/category1/subcategory1/\%$ has been selected, the node $domain/category1/\%$ will be evaluated solely on the URLs not belonging to $domain/category1/subcategory1/\%$). An exterior 'while' loop ensures that we choose the right minimum conversion rate τ_{-} with respect to bounds (r_{min}, r_{max}) on the budget of dimensions of the Hawkes process allowed by our computational architecture.

Backtests on historical data. In order to verify that our model is able to rank accurately the internet user w.r.t. to their probability to convert, we first evaluate it on historical data. To do so, at a given date d (in days), we score each cookie with data collected between $d-90$ and $d-7$, and try to predict the event "a conversion occured between $d-7$ and d". For different segments sizes x, we then plot the conversion rate (i.e. the proportion of cookies having converted between $d-7$ and d) of the x cookies having the highest score according to our Hawkes model and naive selection baselines. In Fig. 2.6, the client operates in the food delivery business, and we compare our model to the baseline consisting in first selecting cookies at random amongst the previous visitors of the website (*retargeting* campaigns) then

Algorithm 3 Preprocessing algorithm for clustering URLs into r classes

input Bounds (r_{min}, r_{max}) for acceptable number of classes, bounds $(\tau_-^{min}, \tau_-^{max})$ for minimum conversion rate, URL tree $\mathscr{T}$ of depth td, minimum number of visits per node vis_-.

```
while |𝒟| ≤ r_min or |𝒟| ≥ r_max do
    𝒟 ← ∅, 𝒰 ← ∅, τ_− ← sqrt(τ_−^min τ_−^max)
    for i ∈ [0...td] do
        n = td − i
        for v ∈ 𝒱_n do
            if vis(v \ 𝒰) ≥ vis_− and τ(v \ 𝒰) ≥ τ_− then
                𝒟 ← 𝒟 ∪ {v}
                𝒰 ← 𝒰 ∪ v
            end if
        end for
    end for
    if |𝒟| ≤ r_min then
        τ_−^max ← τ_−
    end if
    if |𝒟| ≥ r_max then
        τ_−^min ← τ_−
    end if
end while
```

output set $\mathscr{D}$ of URL sets whose visit event constitutes a dimension of the Hawkes process

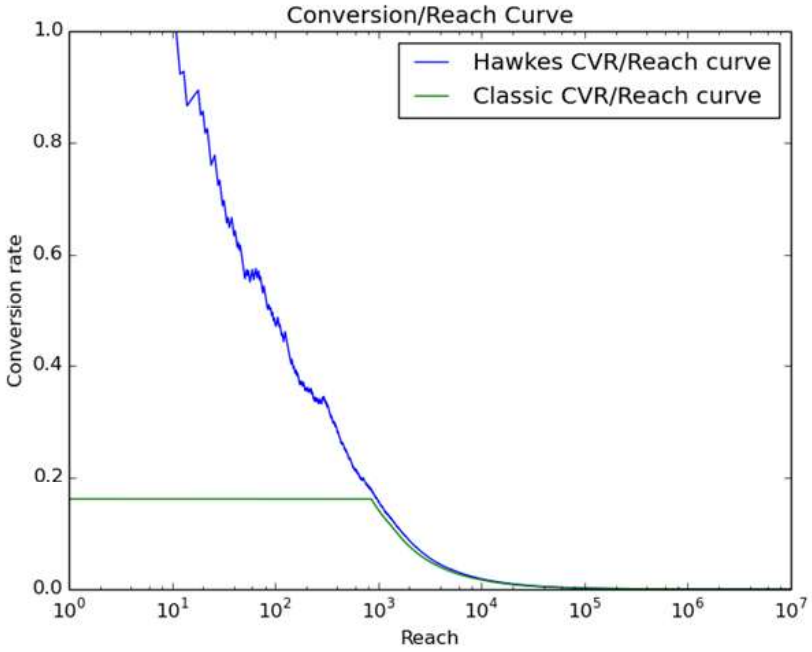


Fig. 2.6 Conversion rate (ordinate) with respect to the size of the segments of top cookies (abscissa) for a food delivery business

selecting cookies at random amongst the rest of the population. We can see that the Hawkes model greatly improves the ability to select the most interesting users. In Fig. 2.7, we focus on the acquisition population, i.e. the population having not previously visited the website. We can observe that the algorithm allows to select very interesting segments. For instance, the top 700000 (resp. 7000000) cookies have a conversion rate multiplied by 8 (resp 2.5) w.r.t. to the average conversion rate of the population. In practice, a segment of size 7000000 cookies is often large enough to spend the whole budget dedicated to the acquisition of new clients. In this example, that would mean acquiring new clients at a cost divided by 2.5 w.r.t. to pure acquisition strategies

Operational results on 1000mercis marketing campaigns. The most relevant way to evaluate our model is to determine whether or not it helped to create successful strategies for 1000mercis marketing campaign. Given an attribution model, a real-time bidding strategy is usually evaluated upon two criterion: the number of conversions it generates, and the average *cost per acquisition* (CPA), i.e. the ratio between the total media cost and the number of leads. However, there are two limits to this evaluation procedure. First, possessing a good user scoring is only an ingredient in the recipe of a successful real-time bidding campaign,

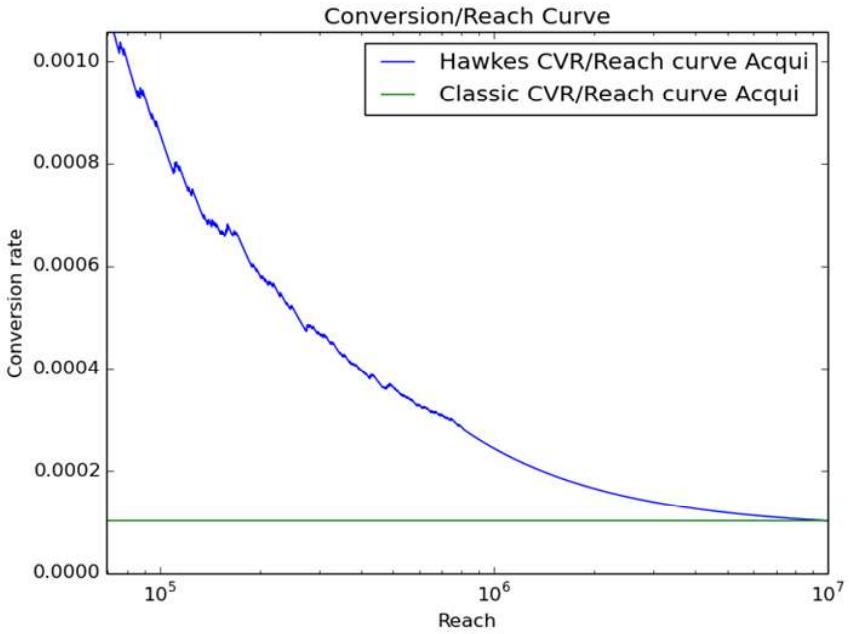


Fig. 2.7 Conversion rate (ordinate) with respect to the size of the segments of top acquisition cookies (abscissa) for a client of the banking sector

and making the most of these strategies necessitates the work of a team of trading experts. We cannot therefore directly evaluate the quality of the scoring by means of the performances of the associated campaigns. Secondly, confidentiality reasons prevent us to develop more than a quick overview of the business impacts. This is however common knowledge that, throughout this PhD, strategies based on internet user scoring by means of Hawkes processes have become a significative part of 1000mercis spendings across several dozens of clients. While each client, and therefore each inferred Hawkes process, is different in terms of the explainability of its conversion and the amount of available data, Hawkes-based strategies have consistently outperformed their counterparts.

2.4 Low-Rank Hawkes Processes (LRHP): an inference algorithm for very large datasets

In this section, we present a way to reduce the inference complexity of any multivariate Hawkes process to $O(d\sum_{h=1}^{H} n_h)$ by introducing *Low-Rank Hawkes Processes* (LRHP), a model for structured point processes relying on a *low-rank decomposition of the triggering kernel* that aim to learn representative patterns of interaction between event types. This

Symbol	Description
d	number of event types, i.e. dimensions of the multivariate Hawkes process
r	rank of the low-dimensional approximation
n	number of events of all realizations of the LRHP process
K	number of triggering kernels
$G=\{\mathscr{V},\mathscr{E}\}$	a network of d nodes, node set $\mathscr{V}$ and edge set $\mathscr{E}$
A	network's adjacency matrix
Δ	maximum node degree of G
$u,v=1,...,d$	indices on dimensions of the original space
$i,j=1,...,r$	indices on dimensions of the low-dimensional embedding
P	$d\times r$ event type-to-group projection matrix
$N(t)=[N_u(t)]_u$	d-dimensional counting process ($t\geq 0, u=1,...,d$)
$\lambda_u(t)$	non-negative occurrence rate for event type u at time t
$\mu_u(t)$	natural occurrence rate for event type u at time t
$g_{vu}(\Delta t)$	kernel function evaluating the affection of λ_u due to events of type v at time distance Δt
α,β	parameters of the triggering kernels
γ,δ	hyperparameters of the triggering kernels
$h=1,...,H$	realizations of the LRHP process (d-dimensional)
$m=1,...,n_h$	events of the realization h, which may belong to any event type
$\mathscr{H}^h$	history of $(t_m^h,u_m^h)_{m=1}^{n_h}$ events of the realization h, indicating (time of event, event type)
$\mathscr{H}$	collection of the event histories of all H realizations
σ	maximum number of event types involved in a realization
B,D	tensors with four and five dimensions, respectively, introduced to simplify our inference algorithm

Table 2.3 Index of main notations.

inference is performed by combining minorize-maximization and self-concordant optimization techniques. We will also assume without loss generality that the interactions between the different dimensions take place along the edges of an unweighted directed network $\mathscr{G}=(\mathscr{V},\mathscr{E})$ of d nodes and adjacency matrix $A\in\{0,1\}^d$. Note that by setting $A_{uv}=1$ for every pair of event types $u\neq v$, we recover the setting of last section where the underlying network of interactions is unknown and each event type can be affected by any other. An index of the main notations used in this section is provided in Tab. 2.3.

2.4.1 Low-Rank Hawkes Processes

Motivations. The main practical issue for inferring the parameters of the model in Eq. 2.9 is that it requires a particularly large dataset of observations, as standard MHP inference requires the learning of d^2 triggering kernels that encode the cross- and self-excitement of the event types. This requirement becomes prohibitive when d is very large (e.g. when the dimensions represent the users of a social network or websites on the Internet). However, in a number of practical situations, the d^2 complex interactions between event types can be summarized by considering that there is a small number of r groups to which each event type belongs to a certain extent. Therefore, one needs to simultaneously learn a $d\times r$ event

type-to-group(s) mapping (we specifically use *soft* assignments) as well as the r^2 interactions between pairs of those groups.

Model formulation. *Low-Rank Hawkes Processes* (LRHP) simplify the standard inference process by projecting the original d event types (i.e. dimensions) of a multivariate Hawkes process into a smaller and more compact r-dimensional space. The natural occurrence rates μ_u and triggering kernels g_{vu} of Eq. 2.1 are then defined via the low-rank approximation:

$$\begin{aligned} \mu_u(t) &= \textstyle\sum_{i=1}^{r} P_{ui}\, \tilde{\mu}_i(t), \\ g_{vu}(t) &= \textstyle\sum_{i,j=1}^{r} P_{ui} P_{vj}\, \tilde{g}_{ji}(t), \end{aligned} \tag{2.15}$$

where u, v are event types, $P \in \mathbb{R}_+^{d\times r}$ is the projection matrix from the original d-dimensional space to the low-dimensional space, and i, j are its component directions. Besides, this projection can be seen as a low-rank approximation of the kernel matrix g since, in matrix notations, $g = P\tilde{g}P^\top$ and $\tilde{g} \in \mathbb{R}_+^{r\times r}$ is a matrix of size $r \ll d$.

Then, the LRHP occurrence rates are formulated as an extension of Eq. 2.9 that uses an embedding of event types in a low-dimensional space:

$$\begin{aligned} \lambda_u(t) = &\sum_{i=1}^{r} P_{ui}\, \tilde{\mu}_i(t) \\ &+ \sum_{m:t_m<t} \sum_{i,j=1}^{r} P_{ui} P_{u_m j} A_{u_m u}\, \tilde{g}_{ji}(t - t_m). \end{aligned} \tag{2.16}$$

Specifically, if the projection of event type u along the dimension i is given by P_{ui}, then the event type u essentially inherits the natural occurrence rate of events of that component $\tilde{\mu}_i$, with multiplicative weight P_{ui}, that is $\sum_{i=1}^{r} P_{ui}\tilde{\mu}_i$. In addition, if the projection of event type v along each dimension j is given by P_{vj}, then v's effect on event type u will be evaluated by $\sum_{i,j=1}^{r} P_{ui}P_{vj}\tilde{g}_{ji}$.

Keeping in mind that $r \ll d$, the proposed low-rank approximation is a simple and straightforward way to: i) impose regularity to the inferred occurrence rates by introducing constraints to the parameters, and ii) reduce the number of parameters. Specifically, the d natural rates and d^2 triggering kernels are reduced to r and r^2, respectively, with the only extra need of inferring the $d \times r$ elements of the matrix P.

Remark on the uniqueness of the projection. Unless any further assumption is made on the projection matrix P or the low-dimensional kernel $\tilde{g}$, the low-rank decomposition of

the triggering kernel $g = P\tilde{g}P^\top$ is not unique. More specifically, any change of basis in the r-dimensional space will not alter the decomposition. Notwithstanding, *uniqueness* is not required in order to perform the prediction task, and therefore we do not address this issue in the present work.

2.4.2 Log-likelihood

General formulation. For $h = 1, ..., H$, let $\mathscr{H}^h = (t_m^h, u_m^h)_{m \leq n_h}$ be the observed i.i.d. realizations sampled from the Hawkes process, and $\mathscr{H} = (\mathscr{H}^h)_{h \leq H}$ the recorded history of events of all realizations. For each realization h, we denote as $[T_-^h, T_+^h]$ the observation period, and u_m^h and t_m^h are respectively the event type and time of occurrence of the m-th event. Then, the log-likelihood of the observations can be written as:

$$\begin{aligned}\mathscr{L}(P, \mathscr{H}; \mu, g) = \sum_{h=1}^{H} \Bigg[\sum_{m=1}^{n_h} \ln \Bigg(\sum_{i=1}^{r} P_{u_m^h i} \, \tilde{\mu}_i(t_m^h) \\ + \sum_{i,j} \sum_{l : t_l^h < t_m^h} P_{u_m^h i} P_{u_l^h j} A_{u_l^h u_m^h} \, \tilde{g}_{ji}(t_m^h - t_l^h) \Bigg) \\ - \sum_{u,i} P_{ui} \int_{T_-^h}^{T_+^h} \tilde{\mu}_i(s) ds \\ - \sum_{u,v,i,j} P_{ui} P_{vj} A_{vu} \int_{T_-^h}^{T_+^h} \tilde{g}_{ji}(s - t_m^h) ds \Bigg].\end{aligned} \tag{2.17}$$

Our objective is to infer the natural rates $\tilde{\mu}_i$ and triggering kernels $\tilde{g}_{ji}$ by means of log-likelihood maximization. From Eq. 2.17, we see that, for arbitrary $\tilde{g}_{ji}$, a single log-likelihood computation already necessitates $O(\sum_{h=1}^{H} n_h^2)$ triggering kernel evaluations. This is intractable when individual realizations can have a number of events of the order 10^7 or 10^8 (e.g. a viral video when modeling information cascades). This issue can be tackled by relying on a convenient K-approximation introduced in Sec. 2.2. Each natural occurrence rate and kernel function are approximated by a sum of K exponential triggering functions:

$$\begin{aligned}\widehat{\mu}_i^K(t) &= \textstyle\sum_{k=0}^{K} \beta_{i,k} \, e^{-k\gamma t}, \\ \widehat{g}_{ji}^K(t) &= \textstyle\sum_{k=1}^{K} \alpha_{ji,k} \, e^{-k\delta t},\end{aligned} \tag{2.18}$$

where $\gamma, \delta > 0$ are fixed hyperparameter values.

Due to the *memoryless property* of exponential functions, this approximation allows for log-likelihood computations with complexity linear in the number of events, i.e. $O(n = \sum_{h=1}^{H} n_h)$. Results of polynomial approximation theory also ensures fast convergence of the optimal $\widehat{\mu}_i^K$ and $\widehat{g}_{ji}^K$ towards the *true* $\tilde{\mu}_i$ and $\tilde{g}_{ji}$ with respect to K. For instance, if $\tilde{g}_{ji}$ is analytic, then $\sup_{t\in[0,T]} |\widehat{g}_{ji}^K(t) - \tilde{g}_{ji}(t)| = O(e^{-K})$ which means that, for smooth enough functions, choosing $K = 10$ already provides a good approximation.

We therefore search the values of parameters α, β that will maximize the approximated log-likelihood, as well as the most probable projection matrix P, conditionally to the realizations of the process, and under the constraint that the approximated natural rates and triggering kernels remain non-negative. At high-level, this is formally expressed as:

$$\begin{aligned} &\underset{(P,\alpha,\beta)}{\arg\max}\ \widehat{\mathscr{L}}(P, \mathscr{H}; \alpha, \beta) \\ &\text{s.t. } \forall i, j, t, K: \ \ \widehat{\mu}_i^K(t) \geq 0 \ \text{ and } \ \widehat{g}_{ji}^K(t) \geq 0. \end{aligned} \tag{2.19}$$

Above, for clarity of notation, we actually reformulate the log-likelihood by introducing $\widehat{\mathscr{L}}$ that makes implicit the dependency of $\mathscr{L}$ in the fixed hyperparameters K, δ, and γ. Note also that limiting K and r to small values can be seen as a form of regularization, although more refined approaches could be considered in case of training with datasets of very limited size.

Simplification with tensor notation. In order to perform inference efficiently, we now reformulate the log-likelihood using very large and sparse tensors. We also introduce the artificial $(r+1)$-th dimension to the embedding space in order to remove linear terms of the equation and store the β parameters as additional dimensions of α. In detail, let $\alpha_{(r+1)i,k} = \beta_{i,k}$, $\alpha_{j(r+1),k} = 0$, and $P_{(d+1)i} = \mathbb{1}_{\{i=r+1\}}$ (note that $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function), also, $\forall u \in \{1, ..., d\}$, $P_{u(r+1)} = 0$. Now, the log-likelihood of the model can be rewritten in the following way:

$$\begin{aligned} \widehat{\mathscr{L}}(P, \mathscr{H}; \alpha) = &\sum_{h,m} \ln \left(\sum_{u,v,i,j,k} P_{ui} P_{vj} \alpha_{ji,k} D_{h,m,u,v,k} \right) \\ &- \sum_{h,u,v,i,j,k} P_{ui} P_{vj} \alpha_{ji,k} B_{h,u,v,k}, \end{aligned} \tag{2.20}$$

where

$$B_{h,u,v,k} = \begin{cases} \sum_{m=1}^{n_h} J_{v,u,m} f_{k\delta}(T_+^h - t_m^h) & \text{if } v \leq d; \\ f_{k\gamma}(T_+^h - T_-^h) & \text{if } v = d+1; \\ 0 & \text{otherwise,} \end{cases} \tag{2.21}$$

Algorithm 4 Inference: high-level description

Input: history of events $\mathscr{H}$; hyperparameters K, γ, δ; initialized projection matrix P and triggering kernel parameters α
Compute D and B *// see Alg. 5*
for $i \in [1...num\_iters]$ **do**
$\alpha = \arg\max_\alpha \widehat{\mathscr{L}}(P, \mathscr{H}; \alpha)$
s.t. $\widehat{\mu}_i^K \geq 0$ and $\widehat{g}_{ji}^K \geq 0,\ i, j = 1, ..., r$
$P = \arg\max_P \widehat{\mathscr{L}}(P, \mathscr{H}; \alpha)$
end for
output P, α

$$D_{h,m,u,v,k} = \begin{cases} \sum_{l=1}^{n_h} I_{h,m,l,u,v}\, e^{-k\delta(t_m^h - t_l^h)} & \text{if } v \leq d; \\ \mathbb{1}_{\{u_m^h = u\}} e^{-k\gamma(t_m^h - T_-^h)} & \text{if } v = d+1; \\ 0 & \text{otherwise,} \end{cases} \tag{2.22}$$

with

$$f_{kx}(t) = \frac{1 - e^{-kxT}}{kx}, \text{ for } x \text{ in } \{\gamma, \delta\},$$

$$J_{v,u,m} = \mathbb{1}_{\{v = u_m^h\}} A_{vu},$$

$$I_{h,m,l,u,v} = \mathbb{1}_{\{u = u_m^h \,\wedge\, v = u_l^h \,\wedge\, t_l^h < t_m^h\}} A_{vu}.$$

What is suggested by the expressions is the possibility to optimize the approximated log-likelihood, according to the different parameters and projection matrices, by first creating two large and sparse tensors B and D with four and five dimensions, respectively.

2.4.3 The inference algorithm

The inference is performed by alternating optimization between the projection matrix P and the Hawkes parameters α. When all others parameters are fixed, the optimization w.r.t. α is performed using self-concordant function optimization with self-concordant barriers. The technical difficulty of this part is due to the need to ensure that non-negativity constraints are respected. For the optimization w.r.t. P, we introduce new optimization techniques based on a minorize-maximization algorithm. Alg. 4 outlines the general scheme of our optimization algorithm. Recall that our basic notation is indexed in Tab. **??**.

Computing B and D tensors. In order for the inference algorithm to be tractable, special attention has to be paid to the computation of B and D tensors. Alg. 5 describes the computation of the sparse tensors $B = (B_{h,u,v,k})$ and $D = (D_{h,m,u,v,k})$. The most expensive

Algorithm 5 Construction of D and B tensors

Initialize $j = 0$
for all h **do**
  Initialize $(C_v^k = \mathbb{1}_{\{v=d+1\}})_{v\geq 0,k\geq 0}$; $t_0^h = T_-^h$; $(B'_{h,u,k} = 0)_{u\geq 0,k\geq 0}$
  $B'_{h,d+1,k} \leftarrow \frac{1-\exp(-k\gamma(T_+^h - T_-^h))}{k\gamma}$
  for all $m \in [1...n_h]$ **do**
    $dt \leftarrow t_m^h - t_{m-1}^h$
    for all k,v s.t $C_v^k > 0$ **do**
      $C_v^k \leftarrow C_v^k \exp(-\mathbb{1}_{\{v>0\}}(k+1)\delta dt - \mathbb{1}_{\{v=0\}}\gamma dt)$
    end for
    for all k **do**
      $D_{h,m,u,v,k} \leftarrow \mathbb{1}_{\{u=u_m\}} \sum_{v\geq 0} A_{u_m v} C_v^k$
      $B'_{h,u_m,k} \leftarrow B'_{h,u_m,k} + \frac{1-\exp(-k\delta(T_+^h - t_m^h))}{k\delta}$
      $C_{u_m}^k \leftarrow C_{u_m}^k + 1$
    end for
    $j \leftarrow j+1$
  end for
  $B_{h,u,v,k} \leftarrow A_{uv} B'_{h,v,k}$
end for
output B, D

operation in this algorithm is the multiplicative update of all C_v^k with the exponential decay $\exp(-(k+\mathbb{1}_{\{v>0\}})\gamma dt)$. Fortunately, this update only has to be performed for every node v that already appeared in the cascade, which are at most $\sigma \leq d$ (by definition). The complexity of this operation is therefore $O(nK\sigma)$. The number of non-zero elements of D and B is $O(nK\min(\Delta,\sigma))$, where Δ is the maximum number of neighbors of a node in the underlying network $\mathscr{G}$. If $\mathscr{G}$ is sparse, which is usually the case for social networks for instance, then $\Delta \ll d$ and therefore $O(nK\Delta) \ll O(nKd)$. Thus, storing and computing B and D is tractable for large dense graphs and for particularly large sparse graphs. Note that, because computing the log-likelihood requires the computation of occurrence rates at each event time, which depends on the occurrences of all preceding events, the linear complexity in the number of events is only possible because of the memoryless property of the decomposition over a basis of exponentials. Otherwise, the respective complexity would have been at least $\Theta(\sum_{h=1}^H n_h^2 K\sigma)$, with $\sum_{h=1}^H n_h^2 \gg n$.

Hawkes parameters optimization. Updating the Hawkes parameters α requires solving the problem:

$$\begin{aligned} \alpha &= \arg\max_{\alpha} \sum_{h,m} \ln\left(c^{hm\top}\alpha\right) - b^\top \alpha \\ &\text{s.t } \widehat{\mu}_i^K \geq 0 \text{ and } \widehat{g}_{ji}^K \geq 0,\ i,j = 1,...,r, \end{aligned} \tag{2.23}$$

where

$$c^{hm}_{ijk} = \sum_{u,v} P_{ui} P_{vj} D_{h,m,u,v,k},$$

$$b_{ijk} = \sum_{u,v,h} P_{ui} P_{vj} B_{h,u,v,k}.$$

For the sake of inference tractability, we relax the non-negativity constraint and only impose it for the observed time differences:

$$\sum_{k=1}^{K} \alpha_{ji,k} D_{h,m,u,v,k} \geq 0. \tag{2.24}$$

Then, we approximate the constrained maximization problem by an unconstrained one, using the concept of *self-concordant barriers* [55]. More specifically, we choose $\varepsilon > 0$ and solve:

$$\alpha = \arg\max_{\alpha} \sum_{h,m} \left(\ln \left(c^{hm\top} \alpha \right) + \varepsilon b(\alpha) \right) - b^{\top} \alpha, \tag{2.25}$$

where

$$b(\alpha)_{hm} = \sum_{i,j,u,v} \ln \left(\sum_{k=1}^{K} \alpha_{ji,k} D_{h,m,u,v,k} \right). \tag{2.26}$$

A feature of the optimization problem in Eq. 2.25 is that it verifies the *self-concordance property*. Self-concordant functions have the advantage of behaving nicely with barrier optimization methods and are among the rare classes of functions for which explicit convergence rates of Newton methods are known [52]. This is the reason why we chose to perform the unconstrained optimization using Newton's method, which requires $O(nKr^2 + K^3 r^6)$ operations. Note that, since we have n events and aim to learn K Hawkes parameters per pair of groups, we have necessarily $Kr^2 \ll n$. If we do not have $K^2 r^4 \ll n$, we can reduce the complexity by using quasi-Newton methods that necessitates only $O(nKr^2 + K^2 r^4) = O(nKr^2)$ operations. The computation of c, b and $b(\alpha)$ requires multiplying sparse matrices of $O(nK\Delta)$ non-zero elements with a full matrix of r columns, which yields a $O(nK\Delta r)$ complexity. Overall, the complexity of the Hawkes parameters optimization is of the order $O(nKr(\Delta + r))$.

Projection matrix optimization. Let p a reshaping of the projection matrix P to a vector (linearized), then p is updated by solving the following maximization procedure:

$$p = \arg\max_{p} \sum_{h,m} \ln \left(p^{\top} \Xi^{hm} p \right) - p^{\top} \Psi p, \tag{2.27}$$

where

$$2\,\Xi^{hm}_{ui,vj} = \sum_k (\alpha_{ji,k} D_{h,m,u,v,k} + \alpha_{ij,k} D_{h,m,v,u,k}),$$

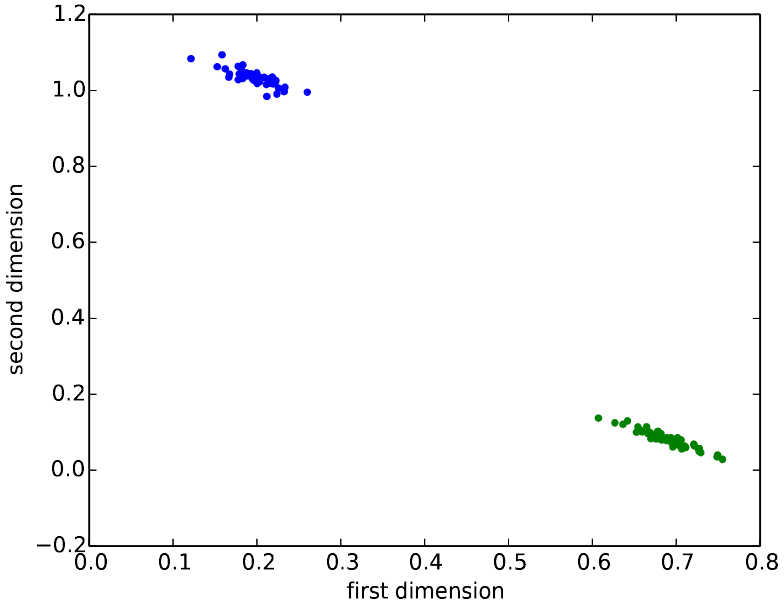


Fig. 2.8 Low-dimensional embedding of the event types learned by LRHP in the synthetic dataset. The two groups (blue and green) of event types are successfully identified.

$$2\Psi_{ui,vj} = \sum_{h,k}(\alpha_{ji,k}B_{h,u,v,k} + \alpha_{ij,k}B_{h,v,u,k}).$$

The maximization task is performed by a novel minorize-maximization procedure which is summarized by the following proposition, proved in Appendix A.

Proposition 9. *The log-likelihood is non-decreasing under the update:*

$$p_{ui}^{t+1} = p_{ui}^{t}\left(\sum_{h,m}\frac{(\Xi^{hm}p^{t})_{ui}}{p^{t\top}\Xi^{hm}p^{t}(\Psi p^{t})_{ui}}\right)^{1/2}. \tag{2.28}$$

Furthermore, if p_{ui} is a stable fixed point of Eq. 2.28, then p_{ui} is a local maximum of the log-likelihood.

As previously, the computation of Ξ, Ψ, and all the matrix-vector products requires $O(nK\Delta r^2)$ operations, and each update necessitates $O(nd)$ operations. Again, we consider settings where we have at least a few events per dimension, so the total complexity of the group affinities optimization is $O(nK\Delta r^2)$.

In total, the complexity of the whole optimization procedure is of the order $O(nK\sigma + nK\Delta r^2)$ and its behavior is linear w.r.t. the number of events and the number of dimensions.

2.4.4 Experimental results

For testing the performance of the proposed LRHP model and the efficiency of our inference algorithm, the experimental study consists of two parts. First, we simulate MHPs on small random networks and verify that the parameters of the simulation are recovered by our algorithm. Second, we provide results on a prediction task for the MemeTracker dataset in order to show that: i) LRHP is highly competitive compared to state-of-the-art inference algorithms on medium-sized datasets, and ii) LRHP is the first framework able to perform large-scale inference for MHPs.

Synthetic data

In this section we illustrate the validity and precision of our method in learning the diffusion parameters of simulated Hawkes processes. More specifically, we simulate MHPs such that event types are separated into two groups of similar activation pattern. In the context of social networks, these groups may encode *influencer-influencee* types of relations. We show that our inference algorithm can recover the groups and corresponding triggering kernels consistently and with high accuracy. Note that LRHP is more generic than this setting, however, we believe that such simple scenario may provide a clearer overview of the capabilities of our approach.

Data generation procedure. The employed procedure for generation of synthetic datasets is as follows. We assume that the MHPs take place on a random Erdös-Rényi [21] network of $d = 100$ event types whose adjacency matrix A is generated with parameter $p = 0.1$ (i.e. 10 neighbors in average). Then, we consider two distinct groups of event types, and assign each event type to one of the groups at random. The natural occurrence rate $\tilde{\mu}_i$ of each group is fixed to a constant value chosen uniformly over $[0, 0.01]$. The triggering kernels between two groups, i and j, are generated as:

$$\tilde{g}_{ij}(t) = \nu_{ij} \frac{\sin\left(\frac{2\pi t}{\omega_{ij}} + \frac{\pi}{2}((i+j) \bmod 2)\right) + 2}{3(t+1)^2}, \tag{2.29}$$

where ω_{ij} and ν_{ij} are sampled uniformly over respectively $[1, 10]$ and $[0, 1/50]$, respectively. These parameter intervals are chosen so that the behavior of the generated process is *non-explosive* [49]. The rationale behind the kernels in Eq. 2.29 is that they present a power-law decreasing intensity that allows long term influence with a periodic behavior. This kind of dynamics could, for instance, represent the daytime cycles of internet users.

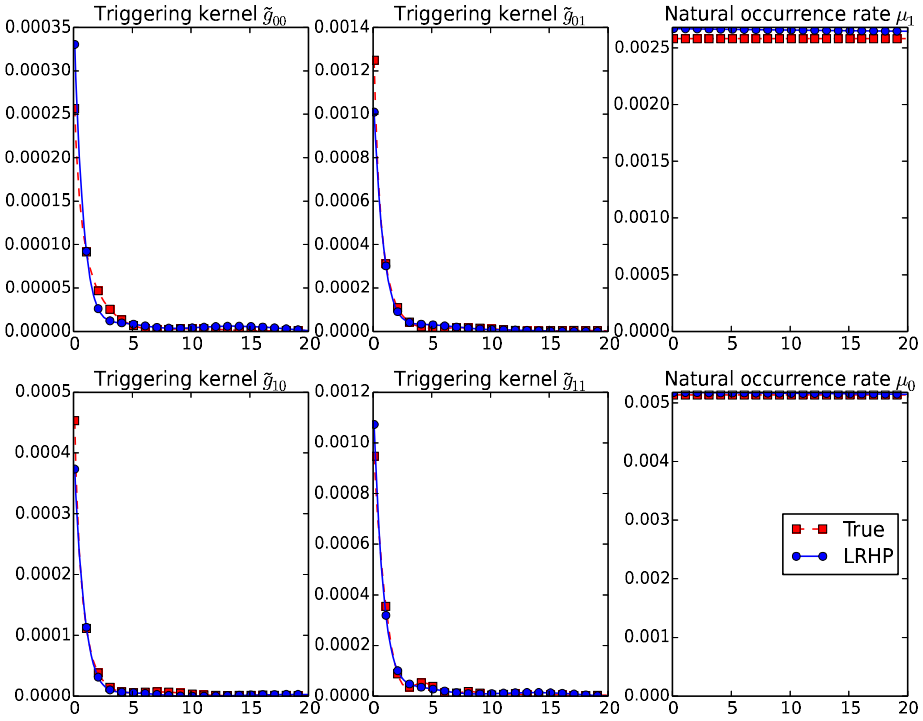


Fig. 2.9 True and inferred triggering kernels $\tilde{g}_{ij}$ and natural occurrence rates $\tilde{\mu}_i$ w.r.t. time (abscissa), for the synthetic dataset.

Results. Following the above procedure we generate 8 datasets by sampling 8 different sets of parameters $\{(\omega_{ij}, \nu_{ij})_{i\leq r, j\leq r}, (\tilde{\mu}_i)_{i\leq r}\}$. Finally, we simulate 10^5 i.i.d. realizations of the resulting Hawkes process, that we use as training set. The ability of LRHP to recover the true group triggering kernels $\tilde{g}_{ij}$, is shown in Fig. 2.9 and evaluated by means of the *normalized* L^2 *error*:

$$\frac{1}{r^2}\sum_{i,j}\frac{||\widehat{g}_{ij}-\tilde{g}_{ij}||_2}{||\widehat{g}_{ij}||_2+||\tilde{g}_{ij}||_2}. \tag{2.30}$$

In average, this is only 12.9%, with minimum and maximum value amongst the 8 datasets of 9.2% and 18.9%, respectively.

In order to find the group assignments, we infer the parameters of an LRHP of rank $r=2$, and recover the group structure by a clustering algorithm on the projected event types. Then, choosing as basis of the two-dimensional space the centers of the two clusters enables the recovery of the group triggering kernels. Fig. 2.8 shows the two-dimensional embedding learned by our inference algorithm for one of the 8 sample datasets. Two particularly separate clusters appear, which indicates that the group assignments were perfectly recovered. The

Table 2.4 Experiments on the MemeTracker datasets. *AUC* (%) and *Accuracy* (%) for predicting the *next event to happen*, using LRHP, MEMIP, and NAIVE approach. In each case, the CPU time (secs) needed for training is also reported. The experiments for the missing measurements, denoted with '$*$', did not finish in reasonable time.

	Dataset			***Training Time***		***AUC***			***Accuracy***		
Name	*thd*	*n*	*d*	LRHP	MEMIP	LRHP	MEMIP	NAIVE	LRHP	MEMIP	NAIVE
MT_1	50000	7311	13	8.34	3.16	86.3	86.4	86.1	99.2	99.2	93.1
MT_2	10000	74474	80	281	$7.14 \cdot 10^3$	90.8	92.6	84.4	89.8	92.7	70.6
MT_3	5000	277914	172	$1.95 \cdot 10^3$	$1.74 \cdot 10^5$	86.2	91.9	81.2	87.0	91.6	67.7
MT_4	1000	875402	1075	$3.77 \cdot 10^5$	*	87.0	*	85.2	84.7	*	81.3

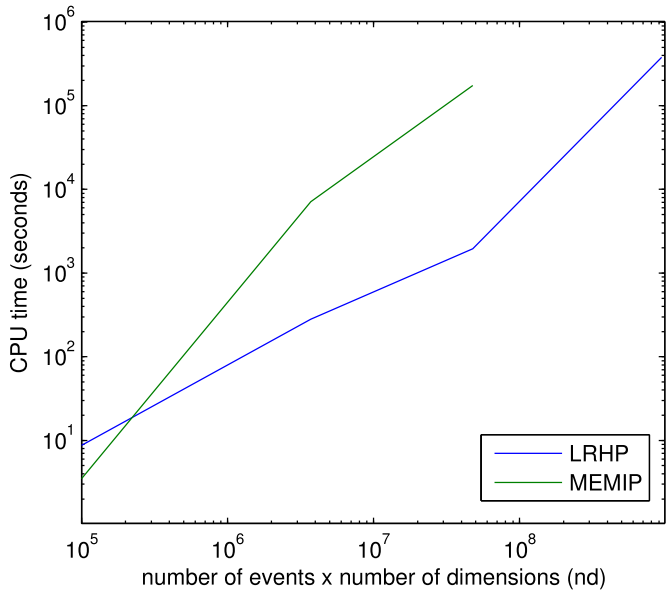


Fig. 2.10 Training time (secs) for LRHP and MEMIP algorithm against the quantity *nd*. The linear behavior for LRHP and super-linear for MEMIP are clearly visible.

other 7 datasets gave similar results. Moreover Fig. 2.9 compares visually the fitness of the inferred to the true natural occurrence rates and triggering kernel functions.

These results provide strong indication regarding the validity of our algorithm for inferring the underlying dynamics of MHPs.

Results on the MemeTracker dataset

Our final set of experiments are conducted on the MemeTracker [61] dataset. MemeTracker is a corpus of $96 \cdot 10^5$ blog posts published between August 2008 and April 2009. We use posts from the period August 2008 to December 2008 as training set, and evaluate our models on the four remaining months. An *event for website u* is defined as the creation of a hyperlink

on website u towards any other website. We also consider that an edge exists between two websites if at least one hyperlink exists between them in the training set. In order to compare the inference algorithms on datasets of different size, prediction was performed on four subsets of the MemeTracker dataset: MT_1, MT_2, MT_3 and MT_4. These subsets are created by removing the events taking place on websites that appear less than a fixed number of times in the training set. This threshold value (*thd* in Tab. 2.4) is, respectively, 50000, 10000, 5000 and 1000.

Prediction task. The task consists in predicting the *next website to create a post*. More specifically, for each event of the test dataset, we are interested in predicting the website on which it will take place knowing its time of occurrence. For MEMIP and LRHP, prediction will be achieved by scoring the websites according to $\lambda_u(t_m)$, since this value is proportional to the theoretical conditional probability for event m to be of type u. We evaluate the prediction with two metrics: the area under the ROC curve (AUC) and a classification accuracy with a fixed number of candidate types. Due to the high bias towards major news websites (e.g. CNN), the number of candidate types has to be relatively large to see differences in the performance of algorithms, and we set this value to 30% of the total number of event types d in our experiments.

Baselines. In the following experiments, we use as main competitor the state-of-the-art MEMIP algorithm [33], which is, to the best of our knowledge, the only inference algorithm with linear complexity in the number of events n in the training history. Also, previous work [33] shows that this algorithm outperforms the more standard inference algorithm MMEL [7] on the MemeTracker dataset. In addition, we also use the NAIVE baseline which ranks the nodes according to their frequency of appearance in the training set. Note that this is equivalent to fitting a Poisson process and, hence, does not consider mutual-excitation.

Results. Tab. 2.4 summarizes the experimental results comparing the proposed LRHP against MEMIP and NAIVE algorithms on four subsets of the MemeTracker dataset. In each row, the table describes the dataset characteristics, and for each method it provides the training time, AUC, and accuracy prediction score. On small to medium-sized datasets (MT_1, MT_2 and MT_3), LRHP is as efficient as its main competitor MEMIP, while orders of magnitude faster. On the large dataset MT_4, LRHP still runs in reasonable time while substantially outperforming the NAIVE baseline. Note that MEMIP could not be computed in reasonable time for this dataset (less than a few days).

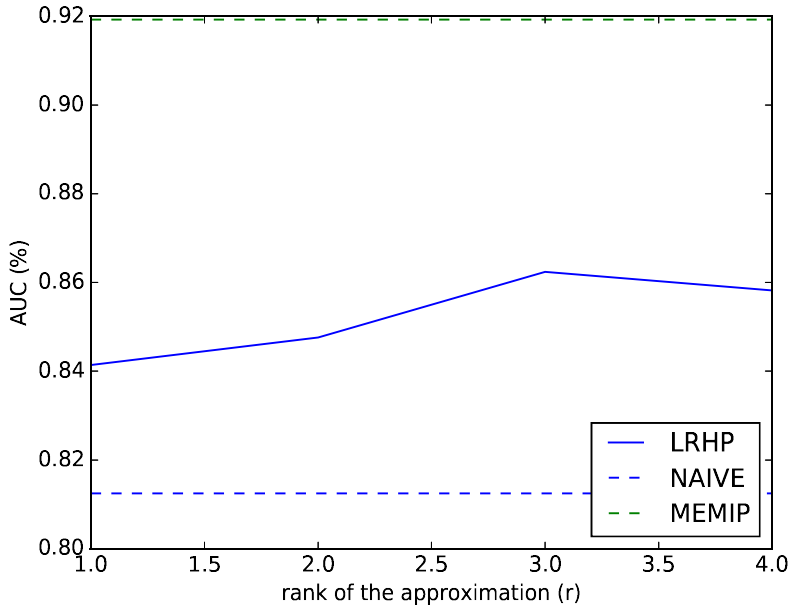


Fig. 2.11 Sensitivity analysis of the AUC of LRHP w.r.t. the rank r of the approximation used for inference, and a comparison to the best scores for MMEL and Naive baselines on the MT_3 dataset.

Fig. 2.10 shows the computational time needed for the inference algorithm on all the MemeTracker datasets, with respect to nd. This time is indeed linear in nd for LRHP, while super-linear for the state-of-the-art competitor of the related literature. In Fig. 2.11 it is indicated that the AUC measurements are relatively stable w.r.t. the rank of the approximation r, with a maximum for $r = 3$. Finally, Fig. 2.12 shows the two-dimensional embedding learned by LRHP for the MT_3 dataset. In the embedding space, the websites seem to align along the axes of the embedding space, with varying amplitudes. This may indicate that two basic groups of similar activities were recovered by the algorithm, although with a large variability in the activity of the websites.

2.4.5 Conclusions

In this chapter, we focus on large scale inference of multivariate Hawkes processes. In the first section, we propose MEMIP, which is to our knowledge the first method to learn nonparametrically triggering kernels of multivariate Hawkes processes in presence of inhibition and varying background rates. By relying on results of approximation theory, the triggering kernels are decomposed on a basis on memoryless exponential kernels. This maximization of the log-likelihood is then shown to reformulate as a concave maximization

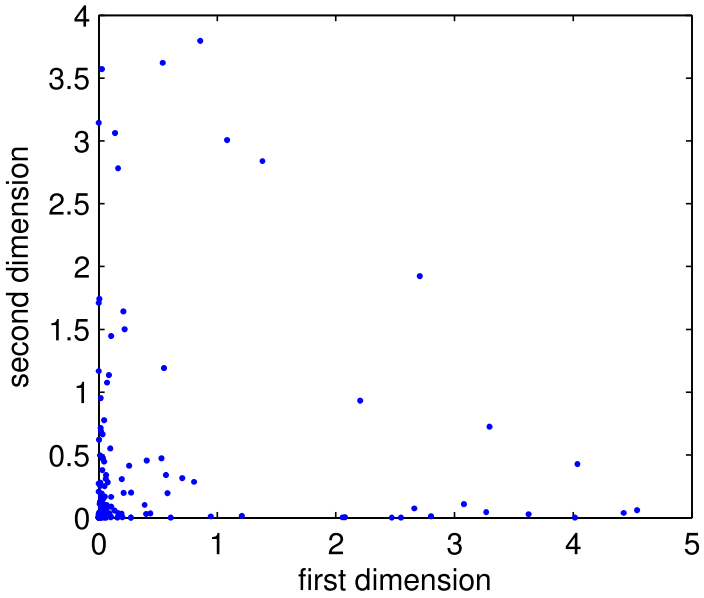


Fig. 2.12 Low-dimensional embeddings of the event types learned by LRHP for the MT_3 dataset.

problem, that can be solved in linear complexity thanks to the Markov property verified by the proposed estimates. Experimental results on both synthetic and real-world data sets show that the proposed model is able to learn more accurately the underlying dynamics of Hawkes processes and therefore has a greater prediction power. In a second part, we focus on modeling multivariate time series where both a very large number of event types can occur, and a very large number of historical observations are available for training. We introduce a model based on multivariate Hawkes processes that we call *Low-Rank Hawkes Processes* (LRHP), and develop an inference algorithm for parameter estimation. Theoretical complexity analysis as well as experimental results show that our approach is highly scalable, while performing as efficiently as state-of-the-art inference algorithms in terms of prediction accuracy.

Chapter 3

Influence bounds

This chapter heavily relies on three papers written with Kevin Scaman and Nicolas Vayatis:

- two of them [62, 63] were published in the conference *Advances in Neural Information Processing Systems* (NIPS) in 2014 and 2015, respectively.
- the third [64] has been submitted to the *Annals of Applied Probability* in 2016.

3.1 Infinite time propagation

3.1.1 Motivations

The emergence of social graphs of the World Wide Web has had a considerable effect on propagation of ideas or information. For advertisers, these new diffusion networks have become a favored vector for *viral marketing* operations, that consist of advertisements that people are likely to share by themselves with their social circle, thus creating a propagation dynamics somewhat similar to the spreading of a virus in epidemiology ([11]). Of particular interest is the problem of *influence maximization*, which consists of selecting the top-k nodes of the network to infect at time $t = 0$ in order to maximize in expectation the final number of infected nodes at the end of the epidemic. This problem was first formulated by Domingues and Richardson in [13] and later expressed in [14] as an NP-hard discrete optimization problem under the Independent Cascade (IC) framework, a widely-used probabilistic model for information propagation.

From an algorithmic point of view, influence maximization has been fairly well studied. Assuming the transmission probability of all edges are known, Kempe, Kleinberg and Tardos ([14]) derived a greedy algorithm based on Monte-Carlo simulations that was shown to approximate the optimal solution up to a factor $1 - \frac{1}{e}$, building on classical results of

optimization theory. Since then, various techniques were proposed in order to significantly improve the scalability of this algorithm ([15–18]), and also to provide an estimate of the transmission probabilities from real data ([19, 20]). Recently, a series of papers ([29, 31, 32]) introduced *continuous-time* diffusion networks in which infection spreads during a time period T at varying rates across the different edges. While these models provide a more accurate representation of real-world networks for finite T, they are equivalent to the IC model when $T \to \infty$. In this section, will focus on such long-term behavior of the contagion.

From a theoretical point of view, little is known about the influence maximization problem under the IC model framework. The most celebrated result established by Newman ([65]) proves the equivalence between bond percolation and the *Susceptible-Infected-Removed* (SIR) model in epidemiology ([26]) that can be identified to a special case of IC model where transmission probability are equal amongst all infectious edges.

In this section, we propose new bounds on the influence of any set of nodes. Moreover, we prove the existence of an *epidemic threshold* for a key quantity defined by the spectral radius of a given *hazard matrix*. Under this threshold, the influence of *any* given set of nodes in a network of size n will be $O(\sqrt{n})$, while the influence of a randomly chosen set of nodes will be $O(1)$. We provide empirical evidence that these bounds are sharp for a family of graphs and sets of initial influencers and can therefore be used as what is to our knowledge the first closed-form formulas for influence estimation. We show that these results generalize bounds obtained on the SIR model by Draief, Ganesh and Massoulié ([27]) and are closely related to recent results on percolation on finite inhomogeneous random graphs ([24]).

The rest of the section is organized as follows. In Sec. 3.1.2, we recall the definition of Information Cascades Model and introduce useful notations. In Sec. 3.1.3, we derive theoretical bounds for the influence. In Sec. 3.1.4, we show that our results also apply to the fields of percolation and epidemiology and generalize existing results in these fields. In Sec. 3.1.5, we illustrate our results by applying them on simple networks and retrieving well-known results. In Sec. 3.1.6, we perform experiments in order to show that our bounds are sharp for a family of graphs and sets of initial nodes.

3.1.2 Information Cascades Model

Influence in random networks and infection dynamics

Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be a directed network of n nodes and $A \subset \mathscr{V}$ be a set of n_0 nodes that are initially *contagious* (e.g. aware of a piece of information, infected by a disease or adopting a product). In the sequel, we will refer to A as the *influencers*. The behavior of the cascade is modeled using a probabilistic framework. The influencer nodes spread the contagion through

the network by means of transmission through the edges of the network. More specifically, each contagious node can infect its neighbors with a certain probability. The *influence* of A, denoted as $\sigma(A)$, is the expected number of nodes reached by the contagion originating from A, i.e.

$$\sigma(A) = \sum_{v \in \mathscr{V}} \mathbb{P}(v \text{ is infected by the contagion } |A). \tag{3.1}$$

We consider three infection dynamics that we will show in the next section to be equivalent regarding the total number of infected nodes at the end of the epidemic.

Discrete-Time Information Cascades [$DTIC(\mathscr{P})$] At time $t = 0$, only the influencers are infected. Given a matrix $\mathscr{P} = (p_{ij})_{ij} \in [0,1]^{n \times n}$, each node i that receives the contagion at time t may transmit it at time $t+1$ along its outgoing edge $(i,j) \in \mathscr{E}$ with probability p_{ij}. Node i cannot make any attempt to infect its neighbors in subsequent rounds. The process terminates when no more infections are possible.

Continuous-Time Information Cascades [$CTIC(\mathscr{F},T)$] At time $t = 0$, only the influencers are infected. Given a matrix $\mathscr{F} = (f_{ij})_{ij}$ of non-negative integrable functions, each node i that receives the contagion at time t may transmit it at time $s > t$ along its outgoing edge $(i,j) \in \mathscr{E}$ with stochastic rate of occurrence $f_{ij}(s-t)$. The process terminates at a given deterministic time $T > 0$. This model is much richer than Discrete-time IC, but we will focus here on its behavior when $T = \infty$.

Random Networks [$RN(\mathscr{P})$] Given a matrix $\mathscr{P} = (p_{ij})_{ij} \in [0,1]^{n \times n}$, each edge $(i,j) \in \mathscr{E}$ is removed independently of the others with probability $1 - p_{ij}$. A node $i \in \mathscr{V}$ is said to be *infected* if i is linked to at least one element of A in the spanning subgraph $\mathscr{G}' = (\mathscr{V}, \mathscr{E}')$ where $\mathscr{E}' \subset \mathscr{E}$ is the set of non-removed edges.

For any $v \in \mathscr{V}$, we will designate by *influence of* v the influence of the set containing only v, i.e. $\sigma(\{v\})$. We will show in Section 3.1.4 that, if $\mathscr{P}$ is symmetric and $\mathscr{G}$ undirected, these three infection processes are equivalent to *bond percolation* and the influence of a node v is also equal to the expected size of the *connected component* containing v in $\mathscr{G}'$. This will make our results applicable to percolation in arbitrary networks. Following the percolation literature, we will denote as *sub-critical* a cascade whose influence is not proportional to the size of the network n.

The hazard matrix

In order to linearize the influence problem and derive upper bounds, we introduce the concept of *hazard matrix*, which describes the behavior of the information cascade. As we will see in the following, in the case of Continuous-time Information Cascades, this matrix gives, for each edge of the network, the integral of the instantaneous rate of transmission (known as hazard function). The spectral radius of this matrix will play a key role in the influence of the cascade.

Definition. For a given graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ and edge transmission probabilities p_{ij}, let $\mathscr{H}$ be the $n \times n$ matrix, denoted as the *hazard matrix*, whose coefficients are

$$\mathscr{H}_{ij} = \begin{cases} -\ln(1-p_{ij}) & \text{if } (i,j) \in \mathscr{E} \\ 0 & \text{otherwise} \end{cases} . \tag{3.2}$$

Next lemma shows the equivalence between the three definitions of the previous section.

Lemma 2. *For a given graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, set of influencers A, and transmission probabilities matrix $\mathscr{P}$, the distribution of the set of infected nodes is equal under the infection dynamics $DTIC(\mathscr{P}), CTIC(\mathscr{F}, \infty)$ and $RN(\mathscr{P})$, provided that for any $(i,j) \in \mathscr{E}$, $\int_0^\infty f_{ij}(t)dt = \mathscr{H}_{ij}$.*

Definition. For a given set of influencers $A \subset \mathscr{V}$, we will denote as $\mathscr{H}(A)$ the hazard matrix except for zeros along the columns whose indices are in A:

$$\mathscr{H}(A)_{ij} = \mathbb{1}_{\{j \notin A\}} \mathscr{H}_{ij}. \tag{3.3}$$

We recall that for any square matrix M, its spectral radius $\rho(M)$ is defined by $\rho(M) = \max_i(|\lambda_i|)$ where $\lambda_1, ..., \lambda_n$ are the (possibly repeated) eigenvalues of matrix M. We will also use that, when M is a real square matrix with positive entries, $\rho(\frac{M+M^\top}{2}) = \sup_X \frac{X^\top M X}{X^\top X}$.

Remark. When the p_{ij} are small, the hazard matrix is very close to the transmission matrix $\mathscr{P}$. This implies that, for low p_{ij} values, the spectral radius of $\mathscr{H}$ will be very close to that of $\mathscr{P}$. More specifically, a simple calculation holds

$$\rho(\mathscr{P}) \leq \rho(\mathscr{H}) \leq \frac{-\ln(1-\|\mathscr{P}\|_\infty)}{\|\mathscr{P}\|_\infty} \rho(\mathscr{P}), \tag{3.4}$$

where $\|\mathscr{P}\|_\infty = \max_{i,j} p_{ij}$. The relatively slow increase of $\frac{-\ln(1-x)}{x}$ for $x \to 1^-$ implies that the behavior of $\rho(\mathscr{P})$ and $\rho(\mathscr{H})$ will be of the same order of magnitude even for high (but lower than 1) values of $\|\mathscr{P}\|_\infty$.

3.1.3 Upper bounds for the influence of a set of nodes

Given $A \subset \mathscr{V}$ the set of influencer nodes and $|A| = n_0 < n$, we derive here two upper bounds for the influence of A. The first bound (Proposition 10) applies to any set of influencers A such that $|A| = n_0$. Intuitively, this result correspond to a best-case scenario (or a worst-case scenario, depending on the viewpoint), since we can target any set of nodes so as to maximize the resulting contagion.

Proposition 10. *Define* $\rho_c(A) = \rho(\frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2})$. *Then, for any A such that* $|A| = n_0 < n$, *denoting by* $\sigma(A)$ *the expected number of nodes reached by the cascade starting from A:*

$$\sigma(A) \leq n_0 + \gamma_1(n - n_0), \tag{3.5}$$

where γ_1 *is the smallest solution in* $[0,1]$ *of the following equation:*

$$\gamma_1 - 1 + \exp\left(-\rho_c(A)\gamma_1 - \frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}\right) = 0. \tag{3.6}$$

Corollary 5. *Under the same assumptions:*

- *if* $\rho_c(A) < 1$, $\quad \sigma(A) \leq n_0 + \sqrt{\frac{\rho_c(A)}{1-\rho_c(A)}}\sqrt{n_0(n-n_0)}$,
- *if* $\rho_c(A) \geq 1$, $\quad \sigma(A) \leq n - (n-n_0)\exp\left(-\rho_c(A) - \frac{2\rho_c(A)}{\sqrt{4n/n_0 - 3} - 1}\right)$.

In particular, when $\rho_c(A) < 1$, $\sigma(A) = O(\sqrt{n})$ *and the regime is sub-critical.*

The second result (Proposition 11) applies in the case where A is drawn from a uniform distribution over the ensemble of sets of n_0 nodes chosen amongst n (denoted as $\mathscr{P}_{n_0}(\mathscr{V})$). This result corresponds to the average-case scenario in a setting where the initial influencer nodes are not known and drawn independently of the transmissions over each edge.

Proposition 11. *Define* $\rho_c = \rho(\frac{\mathscr{H}+\mathscr{H}^\top}{2})$. *Assume the set of influencers A is drawn from a uniform distribution over* $\mathscr{P}_{n_0}(\mathscr{V})$. *Then, denoting by* $\sigma_{uniform}$ *the expected number of nodes reached by the cascade starting from A:*

$$\sigma_{uniform} \leq n_0 + \gamma_2(n - n_0), \tag{3.7}$$

where γ_2 is the unique solution in $[0,1]$ of the following equation:

$$\gamma_2 - 1 + \exp\left(-\rho_c \gamma_2 - \frac{\rho_c n_0}{n - n_0}\right) = 0. \tag{3.8}$$

Corollary 6. *Under the same assumptions:*

- *if $\rho_c < 1$,* $\sigma_{uniform} \leq \dfrac{n_0}{1 - \rho_c}$,
- *if $\rho_c \geq 1$,* $\sigma_{uniform} \leq n - (n - n_0)\exp\left(-\dfrac{\rho_c}{1 - \frac{n_0}{n}}\right)$.

In particular, when $\rho_c < 1$, $\sigma_{uniform} = O(1)$ and the regime is sub-critical.

The difference in the sub-critical regime between $O(\sqrt{n})$ and $O(1)$ for the worst and average case influence is an important feature of our results, and is verified in our experiments (see Sec. 3.1.6). Intuitively, when the network is inhomogeneous and contains highly central nodes (e.g. scale-free networks), there will be a significant difference between specifically targeting the most central nodes and random targeting (which will most probably target a peripheral node).

3.1.4 Application to epidemiology and percolation

Building on the celebrated equivalences between the fields of percolation, epidemiology and influence maximization, we show that our results generalize existing results in these fields.

Susceptible-Infected-Removed (SIR) model in epidemiology

We show here that Proposition 1 further improves results on the SIR model in epidemiology. This widely used model was introduced by Kermac and McKendrick ([26]) in order to model the propagation of a disease in a given population. In this setting, nodes represent individuals, that can be in one of three possible states, susceptible (S), infected (I) or removed (R). At $t = 0$, a subset A of n_0 nodes is infected and the epidemic spreads according to the following evolution. Each infected node transmits the infection along its outgoing edge $(i, j) \in \mathscr{E}$ at stochastic rate of occurrence β and is removed from the graph at stochastic rate of occurrence δ. The process ends for a given $T > 0$. It is straightforward that, if the removed events are not observed, this infection process is equivalent to $CTIC(\mathscr{F}, T)$ where

for any $(i,j) \in \mathscr{E}, f_{ij}(t) = \beta \exp(-\delta t)$. The hazard matrix $\mathscr{H}$ is therefore equal to $\frac{\beta}{\delta}\mathscr{A}$ where $\mathscr{A} = \left(\mathbb{1}_{\{(i,j)\in\mathscr{E}\}}\right)_{ij}$ is the adjacency matrix of the underlying network. Note that, by Lemma 2, our results can be used in order to model the total number of infected nodes in a setting where infection and recovery rates of a given node exhibit a non-exponential behavior. For instance, incubation periods for different individuals generally follow a log-normal distribution [28], which indicates that continuous-time IC with a log-normal rate of removal might be well-suited to model some kind of infections.

It was recently shown by Draief, Ganesh and Massoulié ([27]) that, in the case of undirected networks, and if $\beta\rho(\mathscr{A}) < \delta$,

$$\sigma(A) \leq \frac{\sqrt{nn_0}}{1 - \frac{\beta}{\delta}\rho(\mathscr{A})}. \tag{3.9}$$

This result shows, that, when $\rho(\mathscr{H}) = \frac{\beta}{\delta}\rho(\mathscr{A}) < 1$, the influence of set of nodes A is $O(\sqrt{n})$. We show in the next lemma that this result is a direct consequence of Corollary 5: the condition $\rho_c(\mathscr{A}) < 1$ is weaker than $\rho(\mathscr{H}) < 1$ and, under these conditions, the bound of Corollary 5 is tighter.

Lemma 3. *For any symmetric adjacency matrix $\mathscr{A}$, initial set of influencers A such that $|A| = n_0 < n$, $\delta > 0$ and $\beta < \frac{\delta}{\rho(\mathscr{A})}$, we have simultaneously $\rho_c(A) \leq \frac{\beta}{\delta}\rho(\mathscr{A})$ and*

$$n_0 + \sqrt{\frac{\rho_c(A)}{1-\rho_c(A)}}\sqrt{n_0(n-n_0)} \leq \frac{\sqrt{nn_0}}{1 - \frac{\beta}{\delta}\rho(\mathscr{A})}, \tag{3.10}$$

where the condition $\beta < \frac{\delta}{\rho(\mathscr{A})}$ imposes that the regime is sub-critical.

Moreover, these new bounds capture with more accuracy the behavior of the influence in extreme cases. In the limit $\beta \to 0$, the difference between the two bounds is significant, because Proposition 10 yields $\sigma(A) \to n_0$ whereas (3.9) only ensures $\sigma(A) \leq \sqrt{nn_0}$. When $n = n_0$, Proposition 10 also ensures that $\sigma(A) = n_0$ whereas (3.9) yields $\sigma(A) \leq \frac{n_0}{1-\frac{\beta}{\delta}\rho(\mathscr{A})}$. Secondly, Proposition 10 gives also bounds in the case $\beta\rho(\mathscr{A}) \geq \delta$. Finally, Proposition 10 applies to more general cases that the classical homogeneous SIR model, and allows infection and recovery rates to vary across individuals.

Bond percolation

Given a finite undirected graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, *bond percolation* theory describes the behavior of connected clusters of the spanning subgraph of $\mathscr{G}$ obtained by retaining a subset $\mathscr{E}' \subset \mathscr{E}$

of edges of $\mathscr{G}$ according to a given distribution.When these removals occur independently along each edge with same probability $1-p$, this process is called *homogeneous* percolation and is fairly well known (see e.g [66]). The *inhomogeneous* case, where the independent edge removal probabilities $1-p_{ij}$ vary across the edges, is more intricate and has been the subject of recent studies. In particular, results on critical probabilities and size of the giant component have been obtained by Bollobás, Janson and Riordan in [24]. However, these bounds hold for a particular class of asymptotic graphs (inhomogeneous random graphs) when $n \to \infty$. In the next lemma, we show that our results can be used in order to obtain bounds that hold in expectation for any fixed graph.

Lemma 4. *Let $\mathscr{P} = (p_{ij})_{ij} \in [0,1]^{n\times n}$ be a symmetric matrix. Let $\mathscr{G}' = (\mathscr{V}, \mathscr{E}')$ be the undirected subgraph of $\mathscr{G}$ such that each edge $\{i,j\} \in \mathscr{E}$ is removed independently with probability $1-p_{ij}$. Let $\mathscr{G}_d = (\mathscr{V}, \mathscr{E}_d)$ be the directed graph such that $(i,j) \in \mathscr{E}_d \iff \{i,j\} \in \mathscr{E}$. Then, for any $v \in \mathscr{V}$, the expected size of the connected component containing v in $\mathscr{G}'$ is equal to the influence of v in $\mathscr{G}_d$ under the infection process $DTIC(\mathscr{P})$.*

We now derive an upper bound for $C_1(\mathscr{G}')$, the size of the largest connected component of the spanning subgraph $\mathscr{G}' = (\mathscr{V}, \mathscr{E}')$. In the following, we will denote by $\mathbb{E}[C_1(\mathscr{G}')]$ the expected value of this random variable, given $\mathscr{P} = (p_{ij})_{ij}$.

Proposition 12. *Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be an undirected network where each edge $\{i,j\} \in \mathscr{E}$ has an independent probability $1-p_{ij}$ of being removed. The expected size of the largest connected component of the resulting subgraph $\mathscr{G}'$ is upper bounded by:*

$$\mathbb{E}[C_1(\mathscr{G}')] \leq n\sqrt{\gamma_3}, \tag{3.11}$$

where γ_3 is the unique solution in $[0,1]$ of the following equation:

$$\gamma_3 - 1 + \frac{n-1}{n}\exp\left(-\frac{n}{n-1}\rho(\mathscr{H})\gamma_3\right) = 0. \tag{3.12}$$

Moreover, the resulting network has a probability of being connected upper bounded by:

$$\mathbb{P}(\mathscr{G}' \text{ is connected}) \leq \gamma_3. \tag{3.13}$$

In the case $\rho(\mathscr{H}) < 1$, we can further simplify our bounds in the same way than for Propositions 10 and 11.

Corollary 7. *In the case $\rho(\mathscr{H}) < 1$, $\mathbb{E}[C_1(\mathscr{G}')] \leq \sqrt{\frac{n}{1-\rho(\mathscr{H})}}$.*

Whereas our results hold for any $n \in \mathbb{N}$, classical results in percolation theory study the asymptotic behavior of sequences of graphs when $n \to \infty$. In order to further compare our results, we therefore consider sequences of spanning subgraphs $(\mathscr{G}'_n)_{n\in\mathbb{N}}$, obtained by removing each edge of graphs of n nodes $(\mathscr{G}_n)_{n\in\mathbb{N}}$ with probability $1-p^n_{ij}$. A previous result ([24], Corollary 3.2 of section 5) states that, for particular sequences known as *inhomogeneous random graphs* and under a given sub-criticality condition, $C_1(\mathscr{G}'_n) = o(n)$ *asymptotically almost surely* (a.a.s.), i.e with probability going to 1 as $n \to \infty$. Using Proposition 12, we get for our part the following result:

Corollary 8. *Assume the sequence* $\left(\mathscr{H}^n = \left(-\ln(1-p^n_{ij})\right)_{ij}\right)_{n\in\mathbb{N}}$ *is such that*

$$\limsup_{n\to\infty} \rho(\mathscr{H}^n) < 1. \tag{3.14}$$

Then, for any $\varepsilon > 0$*, we have asymptotically almost surely when* $n \to \infty$*,*

$$C_1(\mathscr{G}'_n) = o(n^{1/2+\varepsilon}). \tag{3.15}$$

This result is to our knowledge the first to bound the expected size of the largest connected component in general arbitrary networks.

3.1.5 Application to particular networks

In order to illustrate our theoretical results, we now apply our bounds to three specific networks and compare them to existing results, showing that our bounds are always of the same order than these specific results. We consider three particular networks: 1) star-shaped networks, 2) Erdös-Rényi networks and 3) random graphs with an expected degree distribution. In order to simplify these problems and exploit existing theorems, we will consider in this section that $p_{ij} = p$ is fixed for each edge $\{i,j\} \in \mathscr{E}$. Infection dynamics thus only depend on p, the set of influencers A, and the structure of the underlying network.

Star-shaped networks

For a star shaped network centered around a given node v_1, and $A = \{v_1\}$, the exact influence is computable and writes $\sigma(\{v_1\}) = 1 + p(n-1)$. As $\mathscr{H}(A)_{ij} = -\ln(1-p)\mathbb{1}_{\{i=1, j\neq 1\}}$, the spectral radius is given by

$$\rho\left(\frac{\mathscr{H}(A) + \mathscr{H}(A)^\top}{2}\right) = \frac{-\ln(1-p)}{2}\sqrt{n-1}. \tag{3.16}$$

Therefore, Proposition 10 states that $\sigma(\{v_1\}) \leq 1+(n-1)\gamma_1$ where γ_1 is the solution of equation

$$1-\gamma_1 = \exp\left(\left(\gamma_1\sqrt{n-1}+\frac{1}{\gamma_1\sqrt{n-1}}\right)\frac{\ln(1-p)}{2}\right). \tag{3.17}$$

It is worth mentionning that, when $p=\frac{1}{\sqrt{n-1}}$, $\gamma_1=\frac{1}{\sqrt{n-1}}$ is solution of (3.17) and therefore the bound is $\sigma(\{v_1\}) \leq 1+\sqrt{n-1}$ which is tight. Note that, in the case of star-shaped networks, the influence does not present a critical behavior and is always linear with respect to the total number of nodes n.

Erdös-Rényi networks

For Erdös-Rényi networks $\mathscr{G}(n,p)$ (*i.e.* an undirected network with n nodes where each couple of nodes $(i,j) \in \mathscr{V}^2$ belongs to $\mathscr{E}$ independently of the others with probability p), the exact influence of a set of nodes is not known. However, percolation theory characterizes the limit behavior of the giant connected component when $n \to \infty$. In the simplest case of Erdös-Rényi networks $\mathscr{G}(n,\frac{c}{n})$ the following result holds:

Lemma 5. *(taken from [24]) For a given sequence of Erdös-Rényi networks $\mathscr{G}(n,\frac{c}{n})$, we have:*

- *if $c<1$, $C_1(\mathscr{G}(n,\frac{c}{n})) \leq \frac{3}{(1-c)^2}\log(n)$ a.a.s.*
- *if $c>1$, $C_1(\mathscr{G}(n,\frac{c}{n})) = (1+o(1))\beta n$ a.a.s. where $\beta-1+\exp(-\beta c)=0$.*

As previously stated, our results hold for any given graph, and not only asymptotically. However, we get an asymptotic behavior consistent with the aforementioned result. Indeed, using notations of section 3.1.4, $\mathscr{H}^n_{ij} = -\ln(1-\frac{c}{n})\mathbb{1}_{\{i\neq j\}}$ and $\rho(\mathscr{H}^n) = -(n-1)\ln(1-\frac{c}{n})$. Using Proposition 12, and noting that $\gamma_3 = (1+o(1))\beta$, we get that, for any $\varepsilon>0$:

- if $c<1$, $C_1(\mathscr{G}(n,\frac{c}{n})) = o(n^{1/2+\varepsilon})$ a.a.s.
- if $c>1$, $C_1(\mathscr{G}(n,\frac{c}{n})) \leq (1+o(1))\beta n^{1+\varepsilon}$ a.a.s., where $\beta-1+\exp(-\beta c)=0$.

Random graphs with given expected degree distribution

In this section, we apply our bounds to random graphs whose expected degree distribution is fixed (see e.g [67], section 13.2.2). More specifically, let $w=(w_i)_{i\in\{1,\dots,n\}}$ be the expected degree of each node of the network. For a fixed w, let $G(w)$ be a random graph whose edges are selected independently and randomly with probability

$$q_{ij} = \frac{\mathbb{1}_{\{i\neq j\}}w_i w_j}{\sum_k w_k}. \tag{3.18}$$

For these graphs, results on the *volume* of connected components (i.e the expected sum of degrees of the nodes in these components) were derived in [68] but our work gives to our knowledge the first result on the size of the giant component. Note that Erdös-Rényi $\mathscr{G}(n,p)$ networks are a special case of (3.18) where $w_i = np$ for any $i \in \mathscr{V}$.

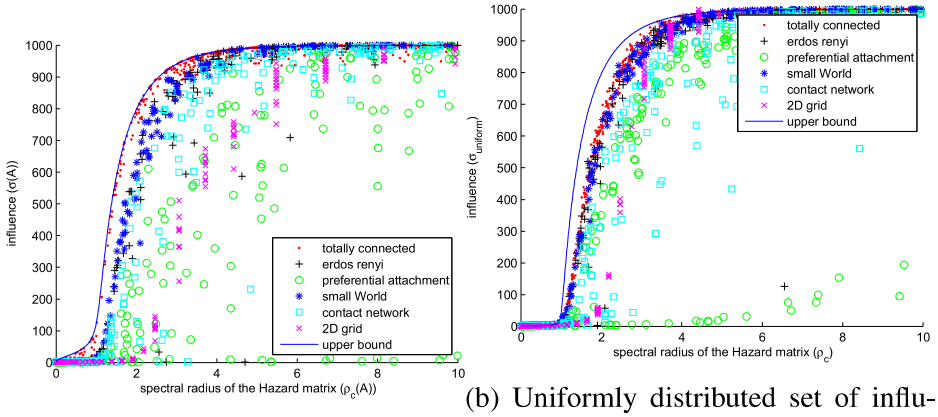
In order to further compare our results, we note that these graphs are also very similar to the widely used *configuration model* where node degrees are fixed to a sequence w, the main difference being that the occupation probabilities p_{ij} are in this case not independent anymore. For configuration models, a giant component exists if and only if $\sum_i w_i^2 > 2\sum_i w_i$ ([69, 70]). In the case of graphs with given expected degree distribution, we retrieve the key role played by the ratio $\sum_i w_i^2 / \sum_i w_i$ in our criterion of non-existence of the giant component given by $\rho(\frac{\mathscr{H}+\mathscr{H}^\top}{2}) < 1$ where

$$\rho\left(\frac{\mathscr{H}+\mathscr{H}^\top}{2}\right) \approx \rho((q_{ij})_{ij}) \leq \frac{\sum_i w_i^2}{\sum_i w_i}. \tag{3.19}$$

The left-hand approximation is particularly good when the q_{ij} are small. This is for instance the case as soon as there exists $\alpha < 1$ such that, for any $i \in \mathscr{V}$, $w_i = o(n^\alpha)$. The right-hand side is based on the fact that the spectral radius of the matrix $(q_{ij} + \mathbb{1}_{\{i=j\}} w_i^2 / \sum_k w_k)_{ij}$ is given by $\sum_i w_i^2 / \sum_i w_i$.

3.1.6 Experimental results

In this section, we show that the bounds given in Sec. 3.1.3 are tight (i.e. very close to empirical results in particular graphs), and are good approximations of the influence on a large set of random networks. Fig. 3.1a compares experimental simulations of the influence to the bound derived in proposition 10. The considered networks have $n = 1000$ nodes and are of 6 types (see e.g [67] for further details on these different networks): 1) Erdös-Rényi networks, 2) Preferential attachment networks, 3) Small-world networks, 4) Geometric random networks ([71]), 5) 2D regular grids and 6) totally connected networks with fixed weight $b \in [0,1]$ except for the ingoing and outgoing edges of the influencer node $A = \{v_1\}$ having weight $a \in [0,1]$. Except for totally connected networks, edge probabilities are set to the same value p for each edge (this parameter was used to tune the spectral radius $\rho_c(A)$). All points of the plots are averages over 100 simulations. The results show that the bound in proposition 10 is tight (see totally connected networks in Fig. 3.1a) and close to the real influence for a large class of random networks. In particular, the tightness of the bound around $\rho_c(A) = 1$ validates the behavior in $\sqrt{n}$ of the worst-case influence in the sub-critical regime. Similarly, Fig. 3.1b compares experimental simulations of the influence to the bound

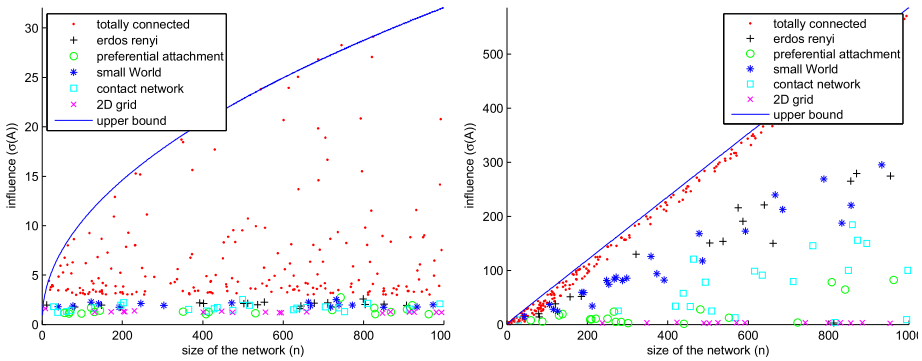


(a) Fixed set of influencers

(b) Uniformly distributed set of influencers

Fig. 3.1 Empirical influence on random networks of various types. The solid lines are the upper bounds in propositions 10 (for Fig. 3.1a) and 11 (for Fig. 3.1b).

derived in proposition 11 in the case of random initial influencers. While this bound is not as tight as the previous one, the behavior of the bound agrees with experimental simulations, and proves a relatively good approximation of the influence under a random set of initial influencers. It is worth mentioning that the bound is tight for the sub-critical regime and shows that corollary 6 is a good approximation of σ_{uniform} when $\rho_c < 1$. In order to verify the criticality of $\rho_c(A) = 1$, we compared the behavior of $\sigma(A)$ w.r.t the size of the network n. When $\rho_c(A) < 1$ (see Fig. 3.2a in which $\rho_c(A) = 0.5$), $\sigma(A) = O(\sqrt{n})$, and the bound is tight. On the contrary, when $\rho_c(A) > 1$ (see Fig. 3.2b in which $\rho_c(A) = 1.5$), $\sigma(A) = O(n)$, and $\sigma(A)$ is linear w.r.t. n for most random networks.



(a) Sub-critical regime: $\rho_c(A) = 0.5$

(b) Super-critical regime: $\rho_c(A) = 1.5$

Fig. 3.2 Influence w.r.t. the size of the network in the sub-critical and super-critical regime. The solid line is the upper bound in proposition 10. Note the square-root versus linear behavior.

3.2 Dynamic Influence Bounds

3.2.1 Motivations

Diffusion networks capture the underlying mechanism of how events propagate throughout a complex network. In marketing, social graph dynamics have caused large transformations in business models, forcing companies to re-imagine their customers not as a mass of isolated economic agents, but as *customer networks* [12]. In epidemiology, a precise understanding of spreading phenomena is heavily needed when trying to break the chain of infection in populations during outbreaks of viral diseases. But whether the subject is a virus spreading across a computer network, an innovative product among early adopters, or a rumor propagating on a network of people, the questions of interest are the same: how many people will it infect? How fast will it spread? And, even more critically for decision makers: how can we modify its course in order to meet specific goals? Several papers tackled these issues by studying the *influence maximization* problem. Given a known diffusion process on a graph, it consists in finding the top-k subset of initial seeds with the highest expected number of infected nodes at a certain time distance T. This problem being NP-hard [14], various heuristics have been proposed in order to obtain scalable suboptimal approximations. While the first algorithms focused on discrete-time models and the special case $T = +\infty$ [15, 16], subsequent papers [29, 30] brought empirical evidences of the key role played by temporal behavior. Existing models of continuous-time stochastic processes include multivariate Hawkes processes [1] where recent progress in inference methods [7, 33] made available the tools for the study of activity shaping [72], which is closely related to influence maximization. However, in the most studied case in which each node of the network can only be infected once, the most widely used model remains the Continuous-Time Information Cascade (CTIC) model [29]. Under this framework, successful inference [29] as well as influence maximization algorithms have been developed [31, 32].

However, if recent works [73, 74] provided theoretical foundations for the inference problem, assessing the quality of influence maximization remains a challenging task, as few theoretical results exist for general graphs. In the infinite-time setting, studies of the SIR diffusion process in epidemiology [27] or percolation for specific graphs [24] provided a more accurate understanding of these processes. We saw in Sec. 3.1 that the spectral radius of a given *Hazard matrix* played a key role in influence of information cascades. This allowed the authors to derive closed-form tight bounds for the influence in general graphs and characterize *epidemic thresholds* under which the influence of any set of nodes is at most $O(\sqrt{n})$.

In this section, we extend their approach in order to deal with the problem of *anytime influence bounds* for continuous-time information cascades. More specifically, we define the *Laplace Hazard matrices* and show that the influence at time T of any set of nodes heavily depends on their spectral radii. Moreover, we reveal the existence and characterize the behavior of *critical times* at which super-critical processes explode. We show that before these times, super-critical processes will behave sub-critically and infect at most $o(n)$ nodes. These results can be used in various ways. First, they provide a way to evaluate influence maximization algorithms without having to test all possible set of influencers, which is intractable for large graphs. Secondly, critical times allow decision makers to know how long a contagion will remain in its early phase before becoming a large-scale event, in fields where knowing *when* to act is nearly as important as knowing *where* to act. Finally, they can be seen as the first closed-form formula for anytime influence estimation for continuous-time information cascades. Indeed, we provide empirical evidence that our bounds are tight for a large family of graphs at the beginning and the end of the infection process.

The rest of the section is organized as follows. In Sec. 3.2.2, we recall the definition of Information Cascades Model and introduce useful notations. In Sec. 3.2.3, we derive theoretical bounds for the influence. In Sec. 3.2.4, we illustrate our results by applying them on specific cascade models. In Sec. 3.2.5, we perform experiments in order to show that our bounds are sharp for a family of graphs and sets of initial nodes. All proof details are provided in the supplementary material.

3.2.2 Continuous-Time Information Cascades

Information propagation and influence in diffusion networks

We describe here the propagation dynamics introduced in [29]. Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be a directed network of n nodes. We equip each directed edge $(i, j) \in \mathscr{E}$ with a time-varying probability distribution $p_{ij}(t)$ over $\mathbb{R}_+ \cup \{+\infty\}$ (p_{ij} is thus a sub-probability measure on $\mathbb{R}_+$) and define the cascade behavior as follows. At time $t = 0$, only a subset $A \subset \mathscr{V}$ of *influencers* is infected. Each node i infected at time τ_i may transmit the infection at time $\tau_i + \tau_{ij}$ along its outgoing edge $(i, j) \in \mathscr{E}$ with probability density $p_{ij}(\tau_{ij})$, and independently of other transmission events. The process ends for a given $T > 0$.

For each node $v \in \mathscr{V}$, we will denote as τ_v the (possibly infinite) time at which it is reached by the infection. The *influence* of A at time T, denoted as $\sigma_A(T)$, is defined as the expected number of nodes reached by the contagion at time T originating from A, i.e.

$$\sigma_A(T) = \mathbb{E}[\sum_{v \in \mathscr{V}} \mathbb{1}_{\{\tau_v \leq T\}}], \tag{3.20}$$

where the expectation is taken over cascades originating from A (i.e. $\tau_v = 0 \Leftrightarrow \mathbb{1}_{\{v \in A\}}$).

Following the percolation literature, we will differentiate between *sub-critical* cascades whose size is $o(n)$ and *super-critical* cascades whose size is proportional to n, where n denotes the size of the network. This work focuses on upper bounding the influence $\sigma_A(T)$ for any given time T and characterizing the critical times at which phase transitions occur between sub-critical and super-critical behaviors.

The Laplace Hazard Matrix

We extend here the concept of *hazard matrix* first introduced in Sec. 3.1, which plays a key role in the influence of the information cascade.

Definition 3. *Let $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ be a directed graph, and p_{ij} be integrable edge transmission probabilities such that $\int_0^{+\infty} p_{ij}(t)dt < 1$. For $s \geq 0$, let $\mathscr{H}(s)$ be the $n \times n$ matrix, denoted as the* Laplace hazard matrix*, whose coefficients are*

$$\mathscr{H}_{ij}(s) = \begin{cases} -\hat{p}_{ij}(s) \left(\int_0^{+\infty} p_{ij}(t)dt \right)^{-1} \ln \left(1 - \int_0^{+\infty} p_{ij}(t)dt \right) & \text{if } (i,j) \in \mathscr{E} \\ 0 & \text{otherwise} \end{cases} . \tag{3.21}$$

where $\hat{p}_{ij}(s)$ denotes the Laplace transform of p_{ij} defined for every $s \geq 0$ by $\hat{p}_{ij}(s) = \int_0^{+\infty} p_{ij}(t)e^{-st}dt$. Note that the long term behavior of the cascade is retrieved when $s = 0$ and coincides with the concept of hazard matrix used in Sec. 3.1.

We recall that for any square matrix M of size n, its spectral radius $\rho(M)$ is the maximum of the absolute values of its eigenvalues. If M is moreover real and positive, we also have $\rho(\frac{M+M^\top}{2}) = \sup_{x \in \mathbb{R}^n} \frac{x^\top M x}{x^\top x}$.

Existence of a critical time of a contagion

In the following, we will derive critical times before which the contagion is sub-critical, and above which the contagion is super-critical. We now formalize this notion of critical time via limits of contagions on networks.

Theorem 2. *Let $(\mathscr{G}_n)_{n \in \mathbb{N}}$ be a sequence of networks of size n, and $(p^n_{ij})_{n \in \mathbb{N}}$ be transmission probability functions along the edges of $\mathscr{G}_n$. Let also $\sigma_n(t)$ be the maximum influence in $\mathscr{G}_n$ at time t from a single influencer. Then there exists a* critical time *$T^c \in \mathbb{R}_+ \cup \{+\infty\}$ such that, for every sequence of times $(T_n)_{n \in \mathbb{N}}$:*

- *If $\limsup_{n \to +\infty} T_n < T^c$, then $\sigma_n(T_n) = o(n)$,*
- *If $\sigma_n(T_n) = o(n)$, then $\liminf_{n \to +\infty} T_n \leq T^c$.*

Moreover, such a critical time is unique.

In other words, the *critical time* is a time before which the regime is *sub-critical* and after which no contagion can be *sub-critical*. The next proposition shows that, after the critical time, the contagion is *super-critical*.

Proposition 13. *If* $(T_n)_{n\in\mathbb{N}}$ *is such that* $\liminf_{n\to+\infty} T_n > T^c$*, then* $\liminf_{n\to+\infty} \frac{\sigma_n(T_n)}{n} > 0$ *and the contagion is* super-critical*. Conversely, if* $(T_n)_{n\in\mathbb{N}}$ *is such that* $\liminf_{n\to+\infty} \frac{\sigma_n(T_n)}{n} > 0$*, then* $\limsup_{n\to+\infty} T_n \geq T^c$*.*

In order to simplify notations, we will omit in the following the dependence in n of all the variables whenever stating results holding in the limit $n \to +\infty$.

3.2.3 Theoretical bounds for the influence of a set of nodes

We now present our upper bounds on the influence at time T and derive a lower bound on the critical time of a contagion.

Upper bounds on the maximum influence at time T

The next proposition provides an upper bound on the influence at time T for any set of influencers A such that $|A| = n_0$. This result may be valuable for assessing the quality of influence maximization algorithms in a given network.

Proposition 14. *Define* $\rho(s) = \rho(\frac{\mathscr{H}(s)+\mathscr{H}(s)^\top}{2})$*. Then, for any A such that* $|A| = n_0 < n$*, denoting by* $\sigma_A(T)$ *the expected number of nodes reached by the cascade starting from A at time T:*

$$\sigma_A(T) \leq n_0 + (n-n_0) \min_{s\geq 0} \gamma(s) e^{sT}. \tag{3.22}$$

where $\gamma(s)$ *is the smallest solution in* $[0,1]$ *of the following equation:*

$$\gamma(s) - 1 + \exp\left(-\rho(s)\gamma(s) - \frac{\rho(s)n_0}{\gamma(s)(n-n_0)}\right) = 0. \tag{3.23}$$

Corollary 9. *Under the same assumptions:*

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \min_{\{s\geq 0 | \rho(s)<1\}} \left(\sqrt{\frac{\rho(s)}{1-\rho(s)}} e^{sT}\right), \tag{3.24}$$

Note that the long-term upper bound in Sec. 3.1 is a corollary of Proposition 14 using $s = 0$. When $\rho(0) < 1$, Corollary 9 with $s = 0$ implies that the regime is sub-critical for all $T \geq 0$. When $\rho(0) \geq 1$, the long-term behavior may be super-critical and the influence may reach linear values in n. However, at a cost growing exponentially with T, it is always possible to choose a s such that $\rho(s) < 1$ and retrieve a $O(\sqrt{n})$ behavior. While the exact optimal parameter s is in general not explicit, two choices of s derive relevant results: either simplifying e^{sT} by choosing $s = 1/T$, or keeping $\gamma(s)$ sub-critical by choosing s s.t. $\rho(s) < 1$. In particular, the following corollary shows that the contagion explodes at most as $e^{\rho^{-1}(1-\varepsilon)T}$ for any $\varepsilon \in [0,1]$.

Corollary 10. *Let $\varepsilon \in [0,1]$ and $\rho(0) \geq 1$. Under the same assumptions:*

$$\sigma_A(T) \leq n_0 + \sqrt{\frac{n_0(n-n_0)}{\varepsilon}} e^{\rho^{-1}(1-\varepsilon)T}. \tag{3.25}$$

Remark. Since this section focuses on bounding $\sigma_A(T)$ for a given $T \geq 0$, all the aforementioned results also hold for $p_{ij}^T(t) = p_{ij}(t)\mathbb{1}_{\{t \leq T\}}$. This is equivalent to integrating everything on $[0,T]$ instead of $\mathbb{R}_+$, i.e. $\mathscr{H}_{ij}(s) = -\ln(1 - \int_0^T p_{ij}(t)dt)(\int_0^T p_{ij}(t)dt)^{-1} \int_0^T p_{ij}(t)e^{-st}dt$. This choice of $\mathscr{H}$ is particularly useful when some edges are transmitting the contagion with probability 1, see for instance the SI epidemic model in Sec. 3.2.4).

Lower bound on the critical time of a contagion

The previous section presents results about how explosive a contagion is. These findings suggest that the speed at which a contagion explodes is bounded by a certain quantity, and thus that the process needs a certain amount of time to become super-critical. This intuition is made formal in the following corollary:

Corollary 11. *Assume $\forall n \geq 0, \rho_n(0) \geq 1$ and $\lim_{n\to+\infty} \frac{\rho_n^{-1}(1-\frac{1}{\ln n})}{\rho_n^{-1}(1)} = 1$. If the sequence $(T_n)_{n\in\mathbb{N}}$ is such that*

$$\limsup_{n\to+\infty} \frac{2\rho_n^{-1}(1)T_n}{\ln n} < 1. \tag{3.26}$$

Then,

$$\sigma_A(T_n) = o(n). \tag{3.27}$$

In other words, the regime of the contagion is *sub-critical* before $\frac{\ln n}{2\rho_n^{-1}(1)}$ and

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\rho_n^{-1}(1)}. \tag{3.28}$$

The technical condition $\lim_{n\to+\infty} \frac{\rho_n^{-1}(1-\frac{1}{\ln n})}{\rho_n^{-1}(1)} = 1$ imposes that, for large n, $\lim_{\varepsilon\to 0} \frac{\rho_n^{-1}(1-\varepsilon)}{\rho_n^{-1}(1)}$ converges sufficiently fast to 1 so that $\rho_n^{-1}(1-\frac{1}{\ln n})$ has the same behavior than $\rho_n^{-1}(1)$. This condition is not very restrictive, and is met for the different case studies considered in Sec. 3.2.4.

This result may be valuable for decision makers since it provides a safe time region in which the contagion has not reached a macroscopic scale. It thus provides insights into *how long* do decision makers have to prepare control measures. After T^c, the process can explode and immediate action is required.

3.2.4 Application to particular contagion models

In this section, we provide several examples of cascade models that show that our theoretical bounds are applicable in a wide range of scenarios and provide the first results of this type in many areas, including two widely used epidemic models.

Fixed transmission pattern

When the transmission probabilities are of the form $p_{ij}(t) = \alpha_{ij} p(t)$ s.t. $\int_0^{+\infty} p(t) = 1$ and $\alpha_{ij} < 1$,

$$\mathscr{H}_{ij}(s) = -\ln(1-\alpha_{ij})\hat{p}(s), \tag{3.29}$$

and

$$\rho(s) = \rho_\alpha \hat{p}(s), \tag{3.30}$$

where $\rho_\alpha = \rho(0) = \rho(-\frac{\ln(1-\alpha_{ij})+\ln(1-\alpha_{ji})}{2})$ is the long-term hazard matrix defined in Sec. 3.1. In these networks, the temporal and structural behaviors are clearly separated. While ρ_α summarizes the structure of the network and how connected the nodes are to one another, $\hat{p}(s)$ captures how fast the transmission probabilities are fading through time.

When $\rho_\alpha \geq 1$, the long-term behavior is super-critical and the bound on the critical times is given by inverting $\hat{p}(s)$

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\hat{p}^{-1}(1/\rho_\alpha)}, \tag{3.31}$$

where $\hat{p}^{-1}(1/\rho_\alpha)$ exists and is unique since $\hat{p}(s)$ is decreasing from 1 to 0. In general, it is not possible to give a more explicit version of the critical time of Corollary 11, or of the anytime influence bound of Proposition 14. However, we investigate in the rest of this section specific $p(t)$ which lead to explicit results.

Exponential transmission probabilities

A notable example of fixed transmission pattern is the case of exponential probabilities $p_{ij}(t) = \alpha_{ij}\lambda e^{-\lambda t}$ for $\lambda > 0$ and $\alpha_{ij} \in [0,1[$. Influence maximization algorithms under this specific choice of transmission functions have been for instance developed in [31]. In such a case, we can calculate the spectral radii explicitly:

$$\rho(s) = \frac{\lambda}{s+\lambda}\rho_\alpha, \tag{3.32}$$

where $\rho_\alpha = \rho(-\frac{\ln(1-\alpha_{ij})+\ln(1-\alpha_{ji})}{2})$ is again the long-term hazard matrix. When $\rho_\alpha > 1$, this leads to a critical time lower bounded by

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\lambda(\rho_\alpha - 1)}. \tag{3.33}$$

The influence bound of Corollary 9 can also be reformulated in the following way:

Corollary 12. *Assume $\rho_\alpha \geq 1$, or else $\lambda T(1-\rho_\alpha) < \frac{1}{2}$. Then the minimum in Eq. 3.24 is met for $s = \frac{1}{2T} + \lambda(\rho_\alpha - 1)$ and Corollary 9 rewrites:*

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)}\sqrt{2eT\lambda\rho_\alpha}e^{\lambda T(\rho_\alpha - 1)}. \tag{3.34}$$

If $\rho_\alpha < 1$ and $\lambda T(1-\rho_\alpha) \geq \frac{1}{2}$, the minimum in Eq. 3.24 is met for $s = 0$ and Corollary 9 rewrites:

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)}\sqrt{\frac{\rho_\alpha}{1-\rho_\alpha}}. \tag{3.35}$$

Note that, in particular, the condition of Corollary 12 is always met in the super-critical case where $\rho_\alpha > 1$. Moreover, we retrieve the $O(\sqrt{n})$ behavior when $T < \frac{1}{\lambda(\rho_\alpha - 1)}$. Concerning the behavior in T, the bound matches exactly the infinite-time bound when T is very large in the sub-critical case. However, for sufficiently small T, we obtain a greatly improved result with a very instructive growth in $O(\sqrt{T})$.

Susceptible-Infected (SI) and Susceptible-Infected-Removed (SIR) epidemic models

Both epidemic models SI and SIR are particular cases of exponential transmission probabilities. SIR model ([26]) is a widely used epidemic model that uses three states to describe the spread of an infection. Each node of the network can be either : susceptible (S), infected (I), or removed (R). At $t = 0$, a subset A of n_0 nodes is infected. Then, each node i infected at time τ_i is removed at an exponentially-distributed time θ_i of parameter δ. Transmission

along its outgoing edge $(i,j) \in \mathscr{E}$ occurs at time $\tau_i + \tau_{ij}$ with conditional probability density $\beta \exp(-\beta \tau_{ij})$, given that node i has not been removed at that time. When the removing events are not observed, SIR is equivalent to *CTIC*, except that transmission along outgoing edges of one node are positively correlated. However, our results still hold in case of such a correlation, as shown in the following result.

Proposition 15. *Assume the propagation follow a SIR model of transmission parameter β and removal parameter δ. Define $p_{ij}(t) = \beta \exp(-(\delta+\beta)t)$ for $(i,j) \in \mathscr{E}$. Let $\mathscr{A} = \left(\mathbb{1}_{\{(i,j)\in\mathscr{E}\}}\right)_{ij}$ be the adjacency matrix of the underlying undirected network. Then, results of Proposition 14 and subsequent corollaries still hold with $\rho(s)$ given by:*

$$\rho(s) = \rho\left(\frac{\mathscr{H}(s) + \mathscr{H}(s)^\top}{2}\right) = \ln\left(1+\frac{\beta}{\delta}\right)\frac{\delta+\beta}{s+\delta+\beta}\rho(\mathscr{A}) \tag{3.36}$$

From this proposition, the same analysis than in the independent transmission events case can be derived, and the critical time for the SIR model is

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2(\delta+\beta)(\ln(1+\frac{\beta}{\delta})\rho(\mathscr{A})-1)}. \tag{3.37}$$

Proposition 16. *Consider the SIR model with transmission rate β, recovery rate δ and adjacency matrix $\mathscr{A}_n$. Assume $\liminf_{n\to+\infty} \ln(1+\frac{\beta}{\delta})\rho(\mathscr{A}_n) > 1$, and the sequence $(T_n)_{n\in\mathbb{N}}$ is such that*

$$\limsup_{n\to+\infty} \frac{2(\delta+\beta)(\ln(1+\frac{\beta}{\delta})\rho(\mathscr{A}_n)-1)T_n}{\ln n} < 1. \tag{3.38}$$

Then,

$$\sigma_A(T_n) = o(n). \tag{3.39}$$

This is a direct corollary of Corollary 11 with $\rho^{-1}(1) = (\delta+\beta)(\ln(1+\frac{\beta}{\delta})\rho(\mathscr{A}_n)-1)$.

The SI model is a simpler model in which individuals of the network remain infected and contagious through time (i.e. $\delta = 0$). Thus, the network is totally infected at the end of the contagion and $\lim_{n\to+\infty} \sigma_A(T) = n$. For this reason, the previous critical time for the more general SIR model is of no use here, and a more precise analysis is required. Following the remark of Sec. 3.2.3, we can integrate p_{ij} on $[0,T]$ instead of $\mathbb{R}_+$, which leads to the following result:

Proposition 17. *Consider the SI model with transmission rate β and adjacency matrix $\mathscr{A}_n$. Assume* $\liminf_{n\to+\infty}\rho(\mathscr{A}_n) > 0$ *and the sequence* $(T_n)_{n\in\mathbb{N}}$ *is such that*

$$\limsup_{n\to+\infty} \frac{\beta T_n}{\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}(1-e^{-\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}})} < 1. \tag{3.40}$$

Then,

$$\sigma_A(T_n) = o(n). \tag{3.41}$$

In other words, the critical time for the SI model is lower bounded by

$$T^c \geq \liminf_{n\to+\infty} \frac{1}{\beta}\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}(1-e^{-\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}}). \tag{3.42}$$

If $\rho(\mathscr{A}_n) = o(\ln n)$ (e.g. for sparse networks with a maximum degree in $O(1)$), the critical time resumes to $T_c \geq \liminf_{n\to+\infty} \frac{1}{\beta}\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}$. However, when the graph is denser and $\rho(\mathscr{A}_n)/\ln n \to +\infty$, then $T_c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\beta\rho(\mathscr{A}_n)}$.

Discrete-time Information Cascade

A final example is the discrete-time contagion in which a node infected at time t makes a unique attempt to infect its neighbors at a time $t+T_0$. This defines the *Information Cascade model*, the discrete-time diffusion model studied by the first works on influence maximization [14, 75, 15, 16]. In this setting, $p_{ij}(t) = \alpha_{ij}\delta_{T_0}(t)$ where δ_{T_0} is the Dirac distribution centered at T_0. The spectral radii are given by

$$\rho(s) = \rho_\alpha e^{-sT_0}, \tag{3.43}$$

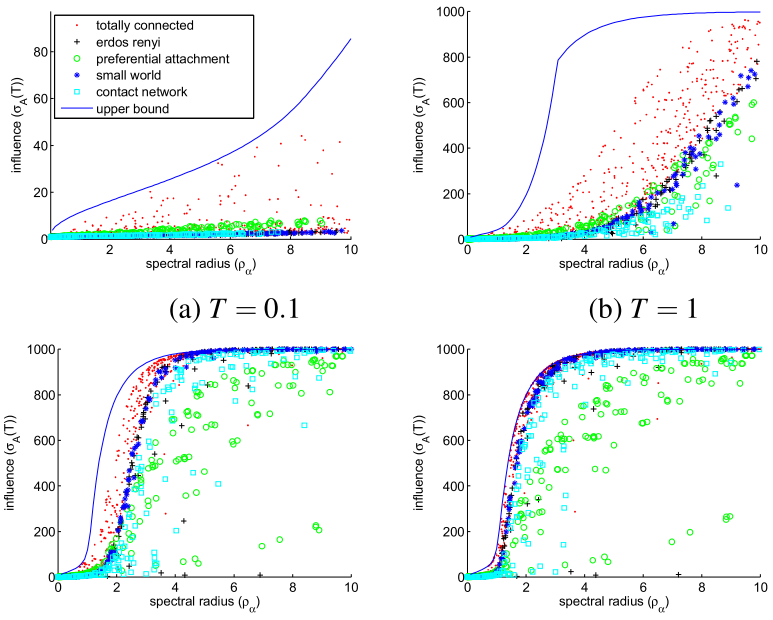
and the influence bound of Corollary 9 simplifies to:

Corollary 13. *Let $\rho_\alpha \geq 1$, or else $T \leq \frac{T_0}{2(1-\rho_\alpha)}$. If $T < T_0$, then $\sigma_A(T) = n_0$. Otherwise,*

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)}\sqrt{\frac{2eT}{T_0}}\rho_\alpha^{\frac{T}{T_0}}. \tag{3.44}$$

Moreover, the critical time is lower bounded by

$$T^c \geq \liminf_{n\to+\infty} \frac{\ln n}{2\ln\rho_\alpha} T_0. \tag{3.45}$$



(a) $T = 0.1$ (b) $T = 1$

(c) $T = 5$ (d) $T = 100$

Fig. 3.3 Empirical maximum influence w.r.t. the spectral radius ρ_α defined in Sec. 3.2.4 for various network types. Simulation parameters: $n = 1000$, $n_0 = 1$ and $\lambda = 1$.

A notable difference from the exponential transmission probabilities is that T^c is here inversely proportional to $\ln \rho_\alpha$, instead of ρ_α in Eq. 3.2.4, which implies that, for the same long-term influence, a discrete-time contagion will explode much slower than one with a constant infection rate. This is probably due to the existence of very small infection times for contagions with exponential transmission probabilities.

3.2.5 Experimental results

This section provides an experimental validation of our bounds, by comparing them to the empirical influence simulated on several network types. In all our experiments, we simulate a contagion with exponential transmission probabilities (see Sec. 3.2.4) on networks of size $n = 1000$ and generated random networks of 5 different types (for more information on the respective random generators, see e.g [67]): Erdös-Rényi networks, preferential attachment networks, small-world networks, geometric random networks ([71]) and totally connected networks with fixed weight $b \in [0, 1]$ except for the ingoing and outgoing edges of a single node having, respectively, weight 0 and $a > b$. The reason for simulating on such totally connected networks is that the influence over these networks tend to match our upper bounds more closely, and plays the role of a best case scenario. More precisely, the transmission

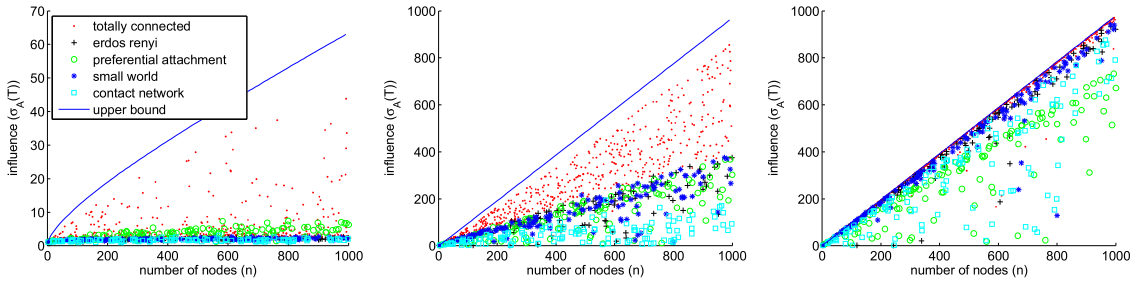
probabilities are of the form $p_{ij}(t) = \alpha e^{-t}$ for each edge $(i,j) \in \mathscr{E}$, where $\alpha \in [0,1[$ (and $\lambda = 1$ in the formulas of Sec. 3.2.4).

We first investigate the tightness of the upper bound on the maximum influence given in Proposition 14. Fig. 3.3 presents the empirical influence w.r.t. $\rho_\alpha = -\ln(1-\alpha)\rho(\mathscr{A})$ (where $\mathscr{A}$ is the adjacency matrix of the network) for a large set of network types, as well as the upper bound in Proposition 14. Each point in the figure corresponds to the maximum influence on one network. The influence was averaged over 100 cascade simulations, and the best influencer (i.e. whose influence was maximal) was found by performing an exhaustive search. Our bounds are tight for all values of $T \in \{0.1, 1, 5, 100\}$ for totally connected networks in the sub-critical regime ($\rho_\alpha < 1$). For the super-critical regime ($\rho_\alpha > 1$), the behavior in T is very instructive. For $T \in \{0.1, 5, 100\}$, we are tight for most network types when ρ_α is high. For $T = 1$ (the average transmission time for the $(\tau_{ij})_{(i,j)\in\mathscr{E}}$), the maximum influence varies a lot across different graphs. This follows the intuition that this is one of the times where, for a given final number of infected node, the local structure of the networks will play the largest role through precise temporal evolution of the infection. Because ρ_α explains quite well the final size of the infection, this discrepancy appears on our graphs at ρ_α fixed. While our bound does not seem tight for this particular time, the order of magnitude of the explosion time is retrieved and our bounds are close to optimal values as soon as $T = 5$.

In order to further validate that our bounds give meaningful insights on the critical time of explosion for super-critical graphs, Fig. 3.4 presents the empirical influence with respect to the size of the network n for different network types and values of T, with ρ_α fixed to $\rho_\alpha = 4$. In this setting, the critical time of Corollary 11 is given by $T^{c*} = \frac{\ln n}{2(\rho_\alpha - 1)\lambda}$. We see that our bounds are tight for totally connected networks for all values of $T \in \{0.2, 2, 5\}$. Moreover, the accuracy of critical time estimation is proved by the drastic change of behavior around $T = T^{c*}$, with phase transitions having occurred for most network types as soon as $T = 5T^{c*}$.

3.2.6 Conclusion

In this chapter, we derived the first upper bounds for the influence of a given set of nodes that apply for any time horizon, graph and set of influencers under the Independent Cascade Model (ICM) framework. We first focus on the long-term influence problem, and relate the influence to the spectral radius of a given *hazard matrix*. We show that these bounds can also be used to generalize previous results in the fields of epidemiology and percolation, and provide empirical evidence that these bounds are close to the best possible for general graphs. In a second part, we generalize these results to continuous-time information cascades for which we characterize the phase transition between sub-critical and super-critical behavior.



(a) $T = 0.2T^{c*}$ (b) $T = 2T^{c*}$ (c) $T = 5T^{c*}$

Fig. 3.4 Empirical maximum influence w.r.t. the network size for various network types. Simulation parameters: $n_0 = 1$, $\lambda = 1$ and $\rho_\alpha = 4$. In such a setting, $T^{c*} = \frac{\ln n}{2(\rho_\alpha - 1)\lambda}$. Note the sub-linear (a) versus linear behavior (b and c).

Indeed, we show that the key quantities governing these continous-time phenomena are the spectral radii of given *Laplace Hazard matrices*, which are built from Laplace transforms of the hazard matrices used for the long term influence. We prove the pertinence of our bounds by deriving the first results of this type in several application fields and we provide experimental evidence that our results are tight for a large family of networks.

Appendix A

Mathematical arguments for Chapter 2

Proof of Proposition 9

For this proof we will make use of the concept of *auxiliary functions*.

Definition 4. *Let g: $\mathscr{X}^2 \to \mathscr{R}$ is an auxiliary function for f: $\mathscr{X} \to \mathscr{R}$ iff $\forall (x,y) \in \mathscr{X}^2, g(x,y) \geq f(x)$ and $\forall x \in \mathscr{X}, g(x,x) = f(x)$.*

The reason why these functions are an important tool for deriving iterative optimization algorithms is given by the following lemma.

Lemma 6. *If g is an auxiliary function for f, then*

$$f\left(\operatorname*{argmin}_x g(x,y)\right) \leq f(y). \tag{A.1}$$

Proof. Let $z = \operatorname{argmin}_x g(x,y)$. Then

$$f(z) = g(z,z) \leq g(z,y) \leq g(y,y) = f(y).$$

where the first inequality comes from the definition of g and the second from the definition of z. □

Therefore, if an auxiliary function g is available, constructing the sequence $y_{t+1} = \operatorname{argmin}_x g(x,y_t)$ that verifies $f(y_{t+1}) \leq f(y_t)$ for all t constitutes a candidate method for finding the minimum of f. In our case, we are able to make use of the following result.

Lemma 7. *Let* $f(p) = -\sum_{k=1}^{K} \ln\left(p^\top \Xi^k p\right) + p^\top \Psi p$ *where* $p \in \mathbb{R}_+^K$, Ξ^1, ..., Ξ^K *are positive symmetric matrices and* Ψ *is a symmetric matrix, then*

$$g(p,q) = -\sum_{k=1}^{K} \left(\frac{2q^\top \Xi^k [q \ln(p/q)]}{q^\top \Xi^k q} + \ln\left(q^\top \Xi^k q\right) \right) + q^\top \Psi [p^2/q] \tag{A.2}$$

is an auxiliary function for f.

In the lemma above, the vectors $[q\ln(p/q)]$ and $[p^2/q]$ are to be understood as coordinate-wise operations, i.e. $(q_i \ln(p_i/q_i))_i$ and $(p_i^2/q_i)_i$.

Proof. It is clear that $g(p,p) = f(p)$ so the proof reduces to showing that $g(p,q) \geq f(p)$. Let $k \leq K$. By concavity of the logarithm function, we have for every weight matrix $(\alpha_{ij})_{ij}$ such that $\sum_{i,j} \alpha_{ij} = 1$,

$$\ln\left(p^\top \Xi^k p\right) \geq \sum_{i,j} \alpha_{ij} \ln\left(\frac{p_i \Xi_{ij}^k p_j}{\alpha_{ij}}\right).$$

Note that the right-hand side term of the equation is well-defined because of the positivity constraint imposed on each Ξ_{ij}^k. By choosing $\alpha_{ij} = q_i \Xi_{ij}^k q_j / q^\top \Xi^k q$, and using the symmetry of Ξ^k, we get:

$$\ln\left(p^\top \Xi^k p\right) \geq \frac{2q^\top \Xi^k [q \ln(p/q)]}{q^\top \Xi^k q} + \ln\left(q^\top \Xi^k q\right).$$

For the right-hand side of the above equation, we use the fact that for every i, j it holds

$$p_i p_j \leq \frac{p_i^2 q_j}{2q_i} + \frac{p_j^2 q_i}{2q_j},$$

and the symmetry of Ψ, in order to conclude that $p^\top \Psi p \leq q^\top \Psi [p^2/q]$. □

Using Lemma 7, we are now in position to prove Proposition 1 by showing that the proposed update p^{t+1} is indeed the global minimum of $g(p, p^t)$. g being the sum of univariate convex functions of the p_i, it is sufficient to show that for every i, the partial derivative of $g(p, p^t)$ with respect to p_i vanishes in p_i^{t+1}. We therefore need:

$$-\sum_k \frac{p_i^t (\Xi^k p^t)_i}{p_i^{t+1} {p^t}^\top \Xi^k p^t} + \frac{p_i^{t+1} (\Psi p^t)_i}{p_i^t} = 0,$$

which only positive solution is given by:

$$p_i^{t+1} = p_i^t \left(\sum_k \frac{(\Xi^k p^t)_i}{{p^t}^\top \Xi^k p^t (\Psi p^t)_i} \right)^{1/2}. \tag{A.3}$$

Finally, if p is a stable fixed point of Eq.3, then, by definition, there exists $\varepsilon > 0$ such that, $\forall p'$ s.t. $||p - p'||_2 \leq \varepsilon$, the iterative algorithm starting at $p^0 = p'$ converges to p. However, since f is continuous, a simple iteration of the inequality of Lemma 6 implies that $f(p') \geq f(p^1) \geq ... \geq \lim_{t \to +\infty} f(p^t) = f(p)$, and p is a local minimum of f.

Appendix B

Mathematical arguments for Chapter 3

Mathematical arguments for Sec. 3.1

Proof of Lemma 2

We prove here the equivalence of propagation dynamics $DTIC(\mathscr{P}), CTIC(\mathscr{F},\infty)$ and $RN(\mathscr{P})$, provided that for any $(i,j)\in\mathscr{E}$, $\int_0^\infty f_{ij}(t)dt = \mathscr{H}_{ij}$. More specifically, we prove the following lemma, that will be useful in the subsequent proofs. In the following, we will denote by X_i the state of node i at the end of the infection process, i.e $X_i = 1$ if the infection has reached node i, and $X_i = 0$ otherwise.

Lemma 8. *Let $\mathscr{G} = (\mathscr{V},\mathscr{E})$ be a given directed network and $A \subset \mathscr{V}$ a set of influencers. For any $i \notin A$, we denote by $\mathscr{Q}_i$ the collection of directed paths (without loops) in $\mathscr{G}$ from A to node i. Then, under the infection processes $DTIC(\mathscr{P}), CTIC(\mathscr{F},\infty)$ and $RN(\mathscr{P})$, we have $\forall i \notin A$,*

$$X_i = 1 - \prod_{q\in\mathscr{Q}_i}(1 - \prod_{(j,l)\in q} E_{jl}), \tag{B.1}$$

where the $(E_{jl})_{jl}$ are independant Bernoulli random variables $E_{jl} \sim \mathscr{B}(p_{jl})$ for infection processes $DTIC(\mathscr{P})$ and $RN(\mathscr{P})$, and $E_{jl} \sim \mathscr{B}\big(1-\exp(-\int_0^\infty f_{jl}(t)dt)\big)$ for infection process $CTIC(\mathscr{F},\infty)$.

Proof. First, note that, for $RN(\mathscr{P})$, the random variables $1_{\{(j,l)\in\mathscr{E}'\}}$ and,for $DTIC(\mathscr{P})$, the indicator function of the events that node j succeeds in infecting node l if j is infected during the process and l is still healthy at that time are independant Bernoulli variables $E_{jl} \sim \mathscr{B}(p_{jl})$ and can all be drawn at $t = 0$. Moreover, by definition of the infection processes, a node $i \in \mathscr{V}$ is reached by the contagion if and only if there exists a path from A to i, such that each

of its edges transmitted the contagion. We thus have for $DTIC(\mathscr{P})$ and $RN(\mathscr{P})$:

$$X_i = 1 - \prod_{q \in \mathscr{Q}_i} (1 - \prod_{(j,l) \in q} E_{jl}). \tag{B.2}$$

For $CTIC(\mathscr{F}, \infty)$, the variables drawn at the beginning of the infection process are the (possibly infinite) times τ_{jl} such that node j will infect node l at time $t_j + \tau_{jl}$ if node j has been infected at time t_j, and node l has not been infected by another node before time $t_j + \tau_{jl}$. By definition, these independent random variables have the following survival function:

$$P(\tau_{jl} < t) = 1 - \exp\left(-\int_0^t f_{jl}(s)ds\right) \tag{B.3}$$

Therefore, we have by the same arguments than previously,

$$X_i = 1 - \prod_{q \in \mathscr{Q}_i} (1 - \prod_{(j,l) \in q} 1_{\{\tau_{jl} < \infty\}}), \tag{B.4}$$

which proves the result for $CTIC(\mathscr{F}, \infty)$, defining $E_{jl} = 1_{\{\tau_{jl} < \infty\}}$ □

Lemma 2 is then a direct corollary of Lemma 8 in the case where, for any $(j,l) \in \mathscr{E}$, $\int_0^\infty f_{jl}(t)dt = \mathscr{H}_{jl}$.

Proofs of Proposition 10 and Corollary 5

We develop here the full proofs for Proposition 10 and Corollary 5 that apply to any set of initially infected nodes. We will first need to prove two useful results: Lemma 9, that proves for $j \in \mathscr{V}$ a positive correlation between the events 'node j did not infect node i during the epidemic' and Lemma 11, that bound the probability that a given node gets infected during the infection process.

Lemma 9. *$\forall i \notin A$, $\{1 - X_j E_{ji}\}_{j \in \mathscr{V}}$ are positively correlated.*

Proof. We will make use of the FKG inequality ([76]):

Lemma 10. *(FKG inequality) Let L be a finite distributive lattice, and μ a nonnegative function on L, such that, for any $(x,y) \in L^2$,*

$$\mu(x \vee y)\mu(x \wedge y) \leq \mu(x)\mu(y) \tag{B.5}$$

Then, for any non-decreasing function f and g on L

$$\left(\sum_{x\in L} f(x)g(x)\right)\left(\sum_{x\in L}\mu(x)\right) \geq \left(\sum_{x\in L} f(x)\mu(x)\right)\left(\sum_{x\in L} g(x)\mu(x)\right) \tag{B.6}$$

For a given set of influencers A, the X_j are deterministic functions of the independent random variables $(E_{ij})_{ij}$. Thus, let $f_{ij}(\{E_{i'j'}\}_{(i',j')}) = 1 - X_j E_{ji}$. In order to apply the FKG inequality, we first need to show that each $f_{ij} : \{0,1\}^{n^2} \to \{0,1\}$ is decreasing with respect to the natural partial order on $\{0,1\}^{n^2}$ (i.e. $X \leq Y$ if $X_i \leq Y_i$ for all i). Let $u \in \{0,1\}^{n^2}$ be a given transmission state of the edges of the network. In order to prove the decreasing behavior of f_{ij}, it is sufficient to show that $f_{ij}(u)$ is decreasing with respect to every $u_{(i,j)}$.

But from Lemma 8, it is obvious that $X_i(u) = 1 - \prod_{q\in\mathscr{Q}_i}(1 - \prod_{(j,l)\in q} u_{(j,l)})$ is increasing with respect to every $u_{(i,j)}$. This implies that $f_{ij}(u) = 1 - X_j(u)u_{(j,i)}$ is decreasing with respect to every $u_{(i,j)}$ and that $f_{ij} : \{0,1\}^{n^2} \to \{0,1\}$ is decreasing with respect to the natural partial order on $\{0,1\}^{n^2}$.

Finally, since we consider a product measure (due to the independence of the E_{ij}) on a product space, we can apply the FKG inequality to $\{1 - X_j E_{ji}\}_{j\in\{1,\dots,N\}}$, and these random variables are positively correlated. □

The next lemma ensures that the variables X_i satisfy an implicit inequation that will be the starting point of the proof of Proposition 10.

Lemma 11. *For any A such that $|A| = n_0 < n$ and for any $i \notin A$, the probability $\mathbb{E}[X_i]$ that node i will be reached by the contagion originating from A verifies:*

$$\mathbb{E}[X_i] \leq 1 - \exp\left(-\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j]\right) \tag{B.7}$$

Proof. We first note that a node is infected if and only if one of its neighbors is infected, and the respective ingoing edge transmitted the contagion. Thus

$$X_i = 0 \Leftrightarrow \forall j \in \{1,\dots,n\}, X_j = 0 \text{ or } E_{ji} = 0, \tag{B.8}$$

which implies the following alternative expression for X_i:

$$1 - X_i = \prod_j (1 - X_j E_{ji}). \tag{B.9}$$

Moreover, the positive correlation of $\{1-X_jE_{ji}\}_{j\in\{1,\dots,N\}}$ implies that

$$\mathbb{E}[\prod_j(1-X_jE_{ji})] \geq \prod_j \mathbb{E}[1-X_jE_{ji}] \tag{B.10}$$

which leads to

$$\begin{aligned}\mathbb{E}[X_i] &\leq 1-\textstyle\prod_j \mathbb{E}[1-X_jE_{ji}]\\ &= 1-\textstyle\prod_j \big(1-\mathbb{E}[X_j]\mathbb{E}[E_{ji}]\big)\\ &= 1-\exp\big(\textstyle\sum_j \ln(1-\mathbb{E}[X_j]\mathbb{E}[E_{ji}])\big)\\ &\leq 1-\exp\big(\textstyle\sum_j \ln(1-\mathbb{E}[E_{ji}])\mathbb{E}[X_j]\big)\\ &= 1-\exp\big(-\textstyle\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j]\big)\end{aligned} \tag{B.11}$$

since we have on the one hand, for any $x\in[0,1]$ and $a<1$, $\ln(1-ax)\geq\ln(1-a)x$, and on the other hand $\mathbb{E}[E_{ji}]=1-\exp(-\mathscr{H}_{ji})$ by definition of $\mathscr{H}$. □

Using Lemma 11, we are now ready to start the proof of Proposition 10.

Proof of Proposition 10. In order to simplify notations, we define $Z_i=\big(\mathbb{E}[X_i]\big)_i$ that we collect in the vector $Z=(Z_i)_{i\in[1\dots n]}$. Using lemma 11 and convexity of exponential function, we have for any $u\in R^n$ such that $\forall i\in A, u_i=0$ and $\forall i\notin A, u_i\geq 0$,

$$u^\top Z \leq |u|_1\left(1-\textstyle\sum_{i=1}^{n-1}\frac{u_i}{|u|_1}\exp(-(\mathscr{H}^\top Z)_i)\right) \leq |u|_1\left(1-\exp\big(-\frac{Z^\top\mathscr{H}u}{|u|_1}\big)\right) \tag{B.12}$$

where $|u|_1=\sum_i|u_i|$ is the L_1-norm of u.

Now taking $u=(1_{i\notin A}Z_i)_i$ and noting that $\forall i\in\{1,\dots,n\},\forall j\in A,\mathscr{H}(A)_{ij}=0$, we have

$$\frac{Z^\top Z-n_0}{|Z|_1-n_0} \leq 1-\exp\left(-\frac{Z^\top\mathscr{H}(A)Z}{|Z|_1-n_0}\right) \leq 1-\exp\left(-\frac{\rho_c(A)(Z^\top Z-n_0)}{|Z|_1-n_0}-\frac{\rho_c(A)n_0}{|Z|_1-n_0}\right) \tag{B.13}$$

where $\rho_c(A)=\rho(\frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2})$. Defining $y=\frac{Z^\top Z-n_0}{|Z|_1-n_0}$ and $z=|Z|_1-n_0=\sigma(A)-n_0$, the aforementioned inequation rewrites

$$y\leq 1-\exp\left(-\rho_c(A)y-\frac{\rho_c(A)n_0}{z}\right) \tag{B.14}$$

But by Cauchy-Schwarz inequality applied to u, $(n-n_0)(Z^\top Z-n_0)\geq(|Z|_1-n_0)^2$, which means that $z\leq y(n-n_0)$. We now consider the equation

$$x-1+\exp\left(-\rho_c(A)x-\frac{\rho_c(A)n_0}{x(n-n_0)}\right)=0 \tag{B.15}$$

Because the function $f:x\to x-1+\exp\big(-\rho_c(A)x+\frac{\rho_c(A)n_0}{x(n-n_0)}\big)$ is continuous, verifies $f(1)>0$ and $\lim_{x\to0^+}f(x)=-1$, equation B.15 admits a solution γ_1 in $]0,1[$.

We then prove by contradiction that $z \leq \gamma_1(n-n_0)$. Let us assume $z > \gamma_1(n-n_0)$. Then $y \leq 1-\exp\big(-\rho_c(A)y-\frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}\big)$. But the function $h: x \to x-1+\exp\big(-\rho_c(A)x+\frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}\big)$ is convex and verifies $h(0)<0$ and $h(\gamma_1)=0$. Therefore, for any $y>\gamma_1$, $0=f(\gamma_1) \leq \frac{\gamma_1}{y}f(y)+(1-\frac{\gamma_1}{y})f(0)$, and therefore $f(y)>0$. Thus, $y \leq \gamma_1$. But $z \leq y(n-n_0) \leq \gamma_1(n-n_0)$ which yields the contradiction. □

Proof of Corollary 5. We distinguish between the cases $\rho_c(A)>1$ and $\rho_c(A) \leq 1$.

Case $\rho_c(A)<1$. Using Eq. B.15 and the fact that $\exp(z) \geq 1+z$, we get $\gamma_1 \leq \rho_c(A)\gamma_1 + \frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}$ which rewrites $\gamma_1 \leq \sqrt{\frac{\rho_c(A)n_0}{(1-\rho_c(A))(n-n_0)}}$ in the case $\rho_c<1$. Therefore,

$$\sigma(A) \leq n_0 + \sqrt{\frac{\rho_c(A)}{1-\rho_c(A)}}\sqrt{n_0(n-n_0)} \tag{B.16}$$

Case $\rho_c(A) \geq 1$. Using Eq. B.15, we get $\gamma_1 - 1 + \exp(-\frac{\rho_c(A)n_0}{\gamma_1(n-n_0)}) \geq 0$, which implies $\gamma_1 \ln(\frac{1}{1-\gamma_1}) \geq \frac{\rho_c(A)n_0}{n-n_0} \geq \frac{n_0}{n-n_0}$. By concavity of the logarithm, we therefore have $\gamma_1^2 \geq \frac{n_0(1-\gamma_1)}{n-n_0}$ which means that $\gamma_1(n-n_0) \geq \frac{n_0(\sqrt{4n/n_0-3}-1)}{2}$. By plugging this lower bound in Eq. B.15, we obtain

$$\sigma(A) \leq n_0 + \left(1-\exp\big(-\rho_c(A)-\frac{2\rho_c(A)}{\sqrt{4n/n_0-3}-1}\big)\right)(n-n_0) \tag{B.17}$$

□

Proofs of Proposition 11 and Corollary 6

In this subsection, we develop the proofs for Proposition 11 and Corollary 6 in the case when the set of initially infected node is drawn from a uniform distribution over $\mathscr{P}_{n_0}(\mathscr{V})$.

We start with an important lemma that will play the same role in the proof of Proposition 11 than Lemma 11 in the proof of Proposition 10.

Lemma 12. *Define $\rho_c = \rho(\frac{\mathscr{H}+\mathscr{H}^\top}{2})$. Assume A is drawn from an uniform distribution over $\mathscr{P}_{n_0}(\mathscr{V})$. Then, for any $i \in \mathscr{V}$, the probability $\mathbb{E}[X_i]$ that node i will be reached by the contagion satisfies the following implicit inequation:*

$$\mathbb{E}[X_i] \leq 1 - \frac{n-n_0}{n}\exp\left(-\frac{n}{n-n_0}\sum_j \mathscr{H}_{ji}\mathbb{E}[X_i]\right) \tag{B.18}$$

Proof.

$$\begin{aligned}
\mathbb{E}[X_i] &= \mathbb{E}[1_{\{i\in A\}}] + \mathbb{E}[1_{\{i\notin A\}}]\mathbb{E}[\mathbb{E}[X_i|A]|i\notin A] \\
&\leq \tfrac{n_0}{n} + \tfrac{n-n_0}{n}\left(1 - \mathbb{E}[\exp\left(-\textstyle\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j|A]\right)|i\notin A]\right) \\
&\leq \tfrac{n_0}{n} + \tfrac{n-n_0}{n}\left(1 - \exp\left(-\mathbb{E}[\textstyle\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j|A]|i\notin A]\right)\right) \\
&= 1 - \tfrac{n-n_0}{n}\exp\left(-\textstyle\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j|i\notin A]\right) \\
&\leq 1 - \tfrac{n-n_0}{n}\exp\left(-\tfrac{n}{n-n_0}\textstyle\sum_j \mathscr{H}_{ji}\mathbb{E}[X_j]\right)
\end{aligned} \tag{B.19}$$

where the first inequality is Lemma 11 and the second one is Jensen inequality for conditional expectations. □

Proof of Proposition 11. We define $Z_i = \left(\mathbb{E}[X_i]\right)_i$ that we collect in the vector $Z = (Z_i)_{i\in[1\ldots n]}$. Then, using Lemma 12, and convexity of exponential function, we have:

$$\tfrac{Z^\top Z}{|Z|_1} \leq \left(1 - \tfrac{n-n_0}{n}\textstyle\sum_{i=1}^n \tfrac{Z_i}{|Z|_1}\exp\left(-\tfrac{n}{n-n_0}(\mathscr{H}^\top Z)_i\right)\right) \leq \left(1 - \tfrac{n-n_0}{n}\exp\left(-\tfrac{n}{n-n_0}\tfrac{Z^\top \mathscr{H} Z}{|Z|_1}\right)\right) \tag{B.20}$$

Now, defining $y = \frac{Z^\top Z}{|Z|_1}$, we have by Cauchy-Schwarz inequality $|Z|_1 \leq ny$ where $y \leq 1 - \frac{n-n_0}{n}\exp\left(-\frac{n}{n-n_0}\rho_c y\right)$. Because function $f : x \to x - 1 + \frac{n-n_0}{n}\exp\left(-\frac{n}{n-n_0}\rho_c y\right)$ is continuous and convex over $]0,1[$, $f(0) < 0$ and $f(1) > 0$, there exists a solution $\gamma \in]0,1[$ of the equation $f(x) = 0$. By the same arguments than in proof of Proposition 10 , we have that, for any $z \in [0,1]$, $f(z) \leq 0 \Rightarrow z \leq \gamma$. This proves the uniqueness of γ as well as the fact that $y \leq \gamma$. Now, defining $\gamma_2 = \frac{n_0}{n} + \frac{n-n_0}{n}\gamma$, we have on the one hand

$$\sigma_{\text{uniform}} \leq n_0 + \gamma_2(n - n_0) \tag{B.21}$$

and on the other hand

$$\gamma_2 - 1 + \exp\left(-\rho_c\gamma_2 - \frac{\rho_c n_0}{n - n_0}\right) = 0 \tag{B.22}$$

which proves the proposition. □

Proof of Corollary 6. In the case $\rho_c < 1$, using Proposition 11 and the fact that $\exp(z) \geq 1 + z$, we get $\gamma_2 \leq \rho_c\gamma_2 + \frac{\rho_c n_0}{n-n_0}$ which rewrites $\gamma_2 \leq \frac{\rho_c n_0}{(1-\rho_c)(n-n_0)}$ in the case $\rho_c < 1$. Therefore,

$$\sigma_{\text{uniform}} \leq n_0\left(1 + \frac{\rho_c}{1-\rho_c}\right) = \frac{n_0}{1-\rho_c} \tag{B.23}$$

The second claim is straightforward from Proposition 11 , using the fact that $\gamma_2 \leq 1$. □

Proofs of Lemma 3, Lemma 4, Proposition 12 and Corollary 8

Proof of Lemma 3. Because matrices $\frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2}$ and $\frac{\beta}{\delta}\mathscr{A}$ are symmetric and verify $0 \leq \frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2} \leq \frac{\beta}{\delta}\mathscr{A} = \mathscr{H}$ where $\leq$ stands for the coefficient-wise inequality, we have $\rho(\frac{\mathscr{H}(A)+\mathscr{H}(A)^\top}{2}) \leq \frac{\beta}{\delta}\rho(\mathscr{A})$ as a direct consequence of the Perron-Frobenius theorem (see e.g [77]). We now introduce the function

$$f : \rho \to n_0 + \sqrt{\frac{\rho}{1-\rho}}\sqrt{n_0(n-n_0)} - \frac{\sqrt{nn_0}}{1-\rho}$$

We have $f(0) < 0$ and $f'(\rho) = \sqrt{n_0(n-n_0)}\frac{\rho}{(1-\rho)^{3/2}} - \sqrt{n_0 n}\frac{1}{(1-\rho)^2} < 0$. Therefore, $f(\rho) < 0$ for any $\rho \in [0,1]$, which proves the Lemma. □

Proof of Lemma 4. First, note that, for bond percolation, the random variables $1_{\{\{j,l\}\in\mathscr{E}'\}}$ are independent Bernoulli variables $F_{\{j,l\}} \sim \mathscr{B}(p_{jl})$. We therefore have, similarly than in the proof of Lemma 8

$$X_i = 1 - \prod_{q\in\mathscr{Q}_i}(1 - \prod_{\{j,l\}\in q} F_{\{j,l\}}). \tag{B.24}$$

where X_i is 1 if node i belongs to the connected component containing the influencer node v, and is 0 otherwise. We then show that, because $\mathscr{P}$ is symmetric, for any infection process $DTIC(\mathscr{P})$ on the directed graph $\mathscr{G}_d$, we can also define independent variables $F'_{\{j,l\}} \sim \mathscr{B}(p_{jl})$ such that the final infection state X'_i of node i is:

$$X'_i = 1 - \prod_{q\in\mathscr{Q}_i}(1 - \prod_{\{j,l\}\in q} F'_{\{j,l\}}), \tag{B.25}$$

which proves that X_i and X'_i have the same probability distribution.

Indeed, the event that node j makes an attempt to infect node l will never occur in the same epidemic than the event that node l makes an attempt to infect node j. Therefore, drawing two variables E_{jl} and E_{lj} at the beginning of each epidemic and letting the dynamic decide which of the two results will be used, or drawing only one variable $F'_{\{j,l\}} \sim \mathscr{B}(p_{jl})$ and using it for each epidemic to decide wether the infection can spread along the edge $\{j,l\}$ or not is strictly equivalent, given that E_{jl} and E_{lj} are independent and have the same distribution. From equations B.24 and B.25, we see that, for any $i \in \mathscr{V}$, the probability that a node i is infected is the same for the two processes. □

Proof of Proposition 12. By proposition 11 applied to the case $n_0 = 1$ with the notation $\gamma_3 = \frac{(n-1)\gamma_2+1}{n}$, we get $\sigma_{\text{uniform}} \leq n\gamma_3$. We then use the fact that, when the influencer node is uniformly randomly drawn on $\mathscr{V}$, it belongs to the largest connected component and therefore

creates an infection of $C_1(\mathscr{G}')$ nodes with probability $\frac{C_1(\mathscr{G}')}{n}$. Therefore, $\mathbb{E}[\frac{C_1(\mathscr{G}')}{n}C_1(\mathscr{G}')] \leq \sigma_{\text{uniform}} \leq n\gamma_3$. But $\mathbb{E}[C_1(\mathscr{G}')^2] \geq \mathbb{E}[C_1(\mathscr{G}')]^2$ which yields $\mathbb{E}[C_1(\mathscr{G}')] \leq n\sqrt{\gamma_3}$. Moreover, denoting as $C_A(\mathscr{G}')$ the size of the connected component containing the influencer node, we have $\sigma_{\text{uniform}} = \mathbb{E}[C_A(\mathscr{G}')] = \sum_i i\mathbb{P}(C_A(\mathscr{G}') = i) \geq n\mathbb{P}(C_A(\mathscr{G}') = n) = n\mathbb{P}(\mathscr{G}' \text{ is connected})$, and therefore $\mathbb{P}(\mathscr{G}' \text{ is connected}) \leq \gamma_3$. □

Proof of Corollary 8. According to Eq.3.14 , there exists $m \in \mathbb{N}$ and $\eta < 1$ such that for any $n \geq m$, $\rho(\mathscr{H}^n) \leq \eta$. Therefore, Corrolary 7 implies $\mathbb{E}[C_1(\mathscr{G}'_n)] \leq \sqrt{\frac{n}{1-\eta}}$. But for any $\delta > 0$, $\mathbb{P}(C_1(\mathscr{G}'_n) > \delta n^{1/2+\varepsilon}) \leq \frac{\mathbb{E}[C_1(\mathscr{G}'_n)]}{\delta n^{1/2+\varepsilon}} = o(1)$ which proves the corollary. □

Mathematical arguments for Sec. 3.2

Critical time definition: proofs of Theorem 2 and Proposition 13

Proof of Theorem 2. Let $S = \{T \in \mathbb{R}_+ | \sigma_n(T) = o(n)\}$. S is an interval containing 0 since $\sigma_n(0) = 0$ and, if $T \in S$, then $\forall T' \leq T$, $\sigma_n(T') \leq \sigma_n(T)$ and $T' \in S$. Thus S is of the form $[0, T^c[$ or $[0, T^c]$, and let $T^c = \sup S$ (where $T^c \in \mathbb{R} \cup \{+\infty\}$).

For all time sequences $(T_n)_{n\in\mathbb{N}}$ such that $\limsup_{n\to+\infty} T_n < T^c$, $\exists T < T^c$ and $n' \geq 0$ s.t. , $\forall n \geq n', T_n \leq T$. Hence, by definition of T^c, $\sigma_n(T_n) \leq \sigma_n(T) = o(n)$.

Conversely, if $\sigma_n(T_n) = o(n)$, then $\liminf_{n\to+\infty} T_n \in S$, and $\liminf_{n\to+\infty} T_n \leq T^c$.

Now let $T^{c\prime}$ verify the two constraints of Theorem 2. The first constraint implies that $\forall T < T^{c\prime}$, $T \in S$ and $T \leq T^c$, which leads to $T^{c\prime} \leq T^c$. Moreover, $\forall T < T^c$, $T \in S$ by definition of T^c, and $T \leq T^{c\prime}$ using the second constraint. As a result, $T^{c\prime} = T^c$ and the critical time is unique. □

Proof of Proposition 13. Let $(T_n)_{n\in\mathbb{N}}$ be such that $\liminf_{n\to+\infty} T_n > T^c$. Then $\exists T > T^c$ and $n' \geq 0$ s.t. $\forall n \geq n'$, $T_n \geq T$. However, $T \notin S$ and $\liminf_{n\to+\infty} \sigma_n(T)/n > 0$, which directly implies that $\liminf_{n\to+\infty} \sigma_n(T_n)/n \geq \liminf_{n\to+\infty} \sigma_n(T)/n > 0$.

Conversely, if $(T_n)_{n\in\mathbb{N}}$ is such that $\liminf_{n\to+\infty} \sigma_n(T_n)/n > 0$, then $\limsup_{n\to+\infty} T_n \notin S$ and $\limsup_{n\to+\infty} T_n \geq T^c$. □

Upper bound on the influence: proofs of Proposition 14 and Corollary 5

Let $\tau_i \in \mathbb{R}_+ \cup \{+\infty\}$ be the infection time of node i, and $\tau_{ij} \in \mathbb{R}_+ \cup \{+\infty\}$ the transmission time from node i to node j. Let $A \subset \mathscr{V}$ be a set of influencers, i.e. nodes that are infected at time 0: $\forall i \in A, \tau_i = 0$. Due to the infection dynamic of CTIC, a node $i \notin A$ is infected when at least one of its neighbors is infected, and the respective ingoing edge transmitted the

contagion. We thus have the following equation relating infection times τ_i and τ_{ji} (see for example [32]): $\forall i \notin A$,

$$\tau_i = \min_{j \in \mathscr{V}} \tau_j + \tau_{ji}. \tag{B.26}$$

Let $X_i(t) = \mathbb{1}_{\{\tau_i < t\}}$ be the infection state of node i at time t. Eq. B.26 implies the following equation: $\forall t > 0$ and $i \notin A$,

$$X_i(t) = 1 - \prod_{j \in \mathscr{V}} \left(1 - \mathbb{1}_{\{\tau_j + \tau_{ji} < t\}}\right). \tag{B.27}$$

We now develop the proofs for Proposition 14 and Corollary 5, which rely on upper bounding the Laplace transform of $\sigma_A(T)$.

Lemma 13. *Define* $\rho(s) = \rho(\frac{\mathscr{L}\mathscr{H}(s) + \mathscr{L}\mathscr{H}(s)^\top}{2})$. *Then, for any A such that* $|A| = n_0 < n$, *denoting by* $\hat{\sigma}_A(s) = \int_0^{+\infty} \sigma_A(t) e^{-st} dt$ *the Laplace transform of the expected number of nodes reached by the cascade starting from A at time T:*

$$s\hat{\sigma}_A(s) \leq n_0 + \gamma(s)(n - n_0), \tag{B.28}$$

where $\gamma(s)$ *is the smallest solution in* $[0,1]$ *of the following equation:*

$$\gamma(s) - 1 + \exp\left(-\rho(s)\gamma(s) - \frac{\rho(s) n_0}{\gamma(s)(n - n_0)}\right) = 0. \tag{B.29}$$

This result requires two intermediate lemmas: Lemma 14, that proves for $i \in \mathscr{V}$ and $t > 0$ a positive correlation between the events '*node j did not infect node i before time t*' and Lemma 15, that bounds the probability that a given node gets infected before t.

Lemma 14. $\forall i \notin A$ *and* $t > 0$, $\{1 - \mathbb{1}_{\{\tau_j + \tau_{ji} < t\}}\}_{j \in \mathscr{V}}$ *are positively correlated.*

Proof. Denoting by $\mathscr{Q}_i$ the collection of directed paths in G from the influencers A to node i, we get the following expression for variables $(\tau_i)_{i \in \mathscr{V}}$ [32]:

$$\tau_i = \min_{q \in \mathscr{Q}_i} \sum_{(j,l) \in q} \tau_{jl} \tag{B.30}$$

Therefore, for all $i \notin A$ and $t > 0$, the functions $f_{ij}(\tau_{kl})_{(k,l) \in \mathscr{E}} = \{1 - \mathbb{1}_{\{\tau_j + \tau_{ji} < t\}}\}_{j \in \mathscr{V}}$ are increasing with the partial order on $(\tau_{kl})_{(k,l) \in \mathscr{E}}$ We will then make once again use of Lemma 10, that we recall here for reader's convenience.

Lemma 10. *(FKG inequality) Let L be a finite distributive lattice, and μ a nonnegative function on L, such that, for any $(x,y) \in L^2$,*

$$\mu(x \vee y)\mu(x \wedge y) \leq \mu(x)\mu(y) \tag{B.31}$$

Then, for any non-decreasing function f and g on L

$$\left(\sum_{x\in L} f(x)g(x)\right)\left(\sum_{x\in L}\mu(x)\right) \geq \left(\sum_{x\in L} f(x)\mu(x)\right)\left(\sum_{x\in L} g(x)\mu(x)\right) \tag{B.32}$$

Due to the independence of $(\tau_{kl})_{(k,l)\in\mathscr{E}}$, the condition in Lemma 10 is met by their joint distribution, which is a product measure on the product space $\mathbb{R}^{\mathscr{E}}$. Lemma 14 is then obtained by applying Lemma 10 to any couple of functions $(f_{ij}, f_{ik})_{(i,j)\in\mathscr{E},(i,k)\in\mathscr{E}}$. More specifically, in our problem setting, L is the set of all $(\tau_{kl})_{(k,l)\in\mathscr{E}}$, $\mu(x) = \prod_{(k,l)\in\mathscr{E}} \mathbb{P}(\tau_{kl} = t_{kl})$ is the joint probability distribution of the τ_{kl} when $x = (t_{kl})_{(k,l)\in\mathscr{E}}$.

□

We then show the following lemma that reveals an implicit inequation satisfied by the X_i.

Lemma 15. *For all $(i,j) \in \mathscr{V}^2$, let p_{ij} be an integrable function such that $\int_0^{+\infty} p_{ij}(t)dt < 1$. For any A such that $|A| = n_0 < n$ and for any $i \notin A$, the probability $\mathbb{E}[X_i(t)]$ that node i will be reached by the contagion originating from A verifies:*

$$\mathbb{E}[X_i(t)] \leq 1 - \exp\left(-\sum_j (\mathscr{L}\mathscr{H}_{ji} * \mathbb{E}[X_j])(t)\right), \tag{B.33}$$

where $(f*g)(t) = \int f(s)g(t-s)ds$ stands for the convolution of f with g and $\mathscr{L}\mathscr{H}_{ji}(t) = \frac{-\ln(1-\int_0^{+\infty} p_{ji}(s)ds)}{\int_0^{+\infty} p_{ji}(s)ds} p_{ji}(t)$.

Proof. Eq. B.27 and the positive correlation of $\{1 - \mathbb{1}_{\{\tau_j+\tau_{ji}<t\}}\}_{j\in\{1,\dots,N\}}$ (Lemma 14) imply that

$$\mathbb{E}[X_i(t)] = 1 - \mathbb{E}[\prod_j (1 - \mathbb{1}_{\{\tau_j+\tau_{ji}<t\}})] \leq 1 - \prod_j \mathbb{E}[1 - \mathbb{1}_{\{\tau_j+\tau_{ji}<t\}}] \tag{B.34}$$

which leads to

$$\begin{aligned} \mathbb{E}[X_i(t)] &\leq 1 - \prod_j \left(1 - \mathbb{E}[\mathbb{1}_{\{\tau_j+\tau_{ji}<t\}}]\right) \\ &= 1 - \prod_j \left(1 - \mathbb{E}[\mathbb{E}[X_j(t-\tau_{ji})|\tau_{ji}]]\right), \\ &= 1 - \prod_j \left(1 - \int_0^{+\infty} \mathbb{E}[X_j(s)] p_{ji}(t-s)ds\right), \end{aligned} \tag{B.35}$$

since $\forall i,j \in \mathscr{V}$, τ_j and τ_{ji} are independent and p_{ji} is the probability density of τ_{ji}. Note that, in our setting, we consider that influencer nodes are infected at time 0, and thus are not infectious before $t = 0$. We then linearize the product in Eq. B.35:

$$\begin{aligned}\mathbb{E}[X_i(t)] &\leq 1 - \exp\left(\sum_j \ln(1 - \int_0^{+\infty} \mathbb{E}[X_j(s)] p_{ji}(t-s)ds)\right) \\ &\leq 1 - \exp\left(\sum_j \frac{\ln(1-\int_0^{+\infty} p_{ji}(s)ds)}{\int_0^{+\infty} p_{ji}(s)ds} \int_0^{+\infty} \mathbb{E}[X_j(s)] p_{ji}(t-s)ds\right) \\ &= 1 - \exp\left(-\sum_j (\mathscr{L}\mathscr{H}_{ji} * \mathbb{E}[X_j])(t)\right),\end{aligned} \tag{B.36}$$

since we have on the one hand, for any $x \in [0,1]$ and $a < 1$, $\ln(1-ax) \geq \ln(1-a)x$ (in Eq. B.36, we chose $a = \int_0^{+\infty} p_{ji}(s)ds$ and $x = \frac{\int_0^{+\infty} \mathbb{E}[X_j(s)] p_{ji}(t-s)ds}{\int_0^{+\infty} p_{ji}(s)ds}$), and on the other hand $\mathscr{L}\mathscr{H}_{ji}(t) = \frac{-\ln(1-\int_0^{+\infty} p_{ji}(s)ds)}{\int_0^{+\infty} p_{ji}(s)ds} p_{ji}(t)$ by definition of $\mathscr{L}\mathscr{H}$. Note that $\frac{-\ln(1-\int_0^{+\infty} p_{ji}(s)ds)}{\int_0^{+\infty} p_{ji}(s)ds}$ is approximately 1 when $\int_0^{+\infty} p_{ji}(s)ds$ is close to 0. □

Proof of Lemma 13. From here, Proposition 13 follow from Lemma 15 in the exact same way than, in Sec. 3.1, the proof of Proposition 10 is deduced from Lemma 11. However, we give here the fully detailed proof for sake of completeness.

Let $f_i(s) = \int_0^{+\infty} \mathbb{E}[X_i(t)] s e^{-st} dt$, then, using Jensen's inequality, $\forall i \notin A$ and $s \geq 0$,

$$f_i(s) \leq 1 - \exp\left(-\sum_j \mathscr{H}_{ji}(s) f_j(s)\right), \tag{B.37}$$

where $\mathscr{H}_{ji}(s) = \int_0^{+\infty} \mathscr{L}\mathscr{H}_{ji}(t) e^{-st} dt$ is the Laplace transform of $\mathscr{L}\mathscr{H}_{ji}$. Note also that $\forall i \in A, f_i(s) = 1$.

For every $i \in [1...n]$, we define $Z_i = \big(f_i(s)\big)_i$ and the vector $Z = (Z_i)_{i\in[1...n]}$. Using lemma 15 and convexity of exponential function, we have for any $u \in R^n$ such that $\forall i \in A, u_i = 0$ and $\forall i \notin A, u_i \geq 0$,

$$u^\top Z \leq |u|_1 \left(1 - \sum_{i=1}^{n-1} \frac{u_i}{|u|_1} \exp(-(\mathscr{H}^\top Z)_i)\right) \leq |u|_1 \left(1 - \exp\big(-\frac{Z^\top \mathscr{H} u}{|u|_1}\big)\right) \tag{B.38}$$

where $|u|_1 = \sum_i |u_i|$ is the L_1-norm of u.

Now taking $u = (1_{i\notin A} Z_i)_i$ and noting that $\forall i, u_i \leq Z_i$, we have

$$\frac{Z^\top Z - n_0}{|Z|_1 - n_0} \leq 1 - \exp\left(-\frac{Z^\top \mathscr{H} Z}{|Z|_1 - n_0}\right) \leq 1 - \exp\left(-\frac{\rho(s)(Z^\top Z - n_0)}{|Z|_1 - n_0} - \frac{\rho(s) n_0}{|Z|_1 - n_0}\right) \tag{B.39}$$

where $\rho(s) = \rho(\frac{\mathscr{H}+\mathscr{H}^\top}{2})$. Defining $y = \frac{Z^\top Z - n_0}{|Z|_1 - n_0}$ and $z = |Z|_1 - n_0 = s\hat{\sigma}_A(s) - n_0$, the inequation above rewrites

$$y \leq 1 - \exp\left(-\rho(s)y - \frac{\rho(s)n_0}{z} \right) \tag{B.40}$$

But by Cauchy-Schwarz inequality applied to u, $(n-n_0)(Z^\top Z - n_0) \geq (|Z|_1 - n_0)^2$, which means that $z \leq y(n-n_0)$. We now consider the equation

$$x - 1 + \exp\left(-\rho(s)x - \frac{\rho(s)n_0}{x(n-n_0)} \right) = 0 \tag{B.41}$$

Because the function $f : x \to x - 1 + \exp\big(-\rho(s)x + \frac{\rho(s)n_0}{x(n-n_0)} \big)$ is continuous, verifies $f(1) > 0$ and $\lim_{x\to 0^+} f(x) = -1$, equation B.41 admits a solution $\gamma(s)$ in $]0,1[$.

We then prove by contradiction that $z \leq \gamma(s)(n-n_0)$. Let us assume $z > \gamma(s)(n-n_0)$. Then $y \leq 1 - \exp\big(-\rho(s)y - \frac{\rho(s)n_0}{\gamma(s)(n-n_0)} \big)$. But the function $h : x \to x - 1 + \exp\big(-\rho(s)x + \frac{\rho(s)n_0}{\gamma(s)(n-n_0)} \big)$ is convex and verifies $h(0) < 0$ and $h(\gamma(s)) = 0$. Therefore, for any $y > \gamma_1$, $0 = f(\gamma_1) \leq \frac{\gamma(s)}{y} f(y) + (1 - \frac{\gamma(s)}{y}) f(0)$, and therefore $f(y) > 0$. Thus, $y \leq \gamma(s)$. But $z \leq y(n-n_0) \leq \gamma(s)(n-n_0)$ which yields the contradiction. □

Using Lemma 13, we may now prove Proposition 14:

Proof of Proposition 14. $\forall s \geq 0$, $T \geq 0$ and $t \geq 0$, $e^{-st} \geq e^{-sT}\mathbb{1}_{\{t<T\}}$, hence, using Lemma 13, $s\hat{\sigma}_A(s) = \sum_i \mathbb{E}[e^{-s\tau_i}] \geq n_0 + (\sigma_A(T) - n_0)e^{-sT}$ which leads to the desired inequality. □

Proof of Corollary 5. Using Eq. B.41 and the fact that $1 - e^{-x} \leq x$, we get $\gamma(s) \leq \rho(s)\gamma(s) + \frac{\rho(s)n_0}{\gamma(s)(n-n_0)}$ which rewrites $\gamma(s) \leq \sqrt{\frac{\rho(s)n_0}{(1-\rho(s))(n-n_0)}}$ in the case $\rho(s) < 1$. Therefore,

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \min_{\{s\geq 0 | \rho(s)<1\}} \left(\sqrt{\frac{\rho(s)}{1-\rho(s)}} e^{sT} \right). \tag{B.42}$$

□

Upper bounds on the critical time: proofs of Corollary 10 and Corollary 11

Proof of Corollary 10. Since e^{-st} is decreasing w.r.t. s, $\mathscr{H}_{ij}(s)$ is decreasing. Thus, the Perron-Frobenius theorem implies that $\rho(s)$ is decreasing. When $\rho(0) \geq 1$, $\rho^{-1}(1-\varepsilon)$ exists and is uniquely defined, and using Corollary 5 and 14, $\sigma_A(T) \leq n_0 + (n-n_0)\gamma(\rho^{-1}(1-\varepsilon))e^{\rho^{-1}(1-\varepsilon)T} \leq n_0 + \sqrt{\frac{n_0(n-n_0)}{\varepsilon}} e^{\rho^{-1}(1-\varepsilon)T}$. □

Proof of Corollary 11. If $\limsup_{n\to+\infty} \frac{2\rho^{-1}(1)T_n}{\ln n} < 1$, then $\exists \alpha > 0$ and $n' \geq 0$ s.t. $\forall n \geq n'$, $\rho^{-1}(1)T_n \leq \frac{(1-\alpha)\ln n}{2}$. Furthermore, $\lim_{n\to+\infty} \frac{\rho^{-1}(1-\frac{1}{\ln n})}{\rho^{-1}(1)} = 1$, thus $\exists n'' \geq n'$ s.t. $\forall n \geq n''$, $\rho^{-1}(1-\frac{1}{\ln n}) \leq \frac{1-\alpha/2}{1-\alpha}\rho^{-1}(1)$. Using Corollary 10 with $\varepsilon = \frac{1}{\ln n}$, $\sigma_A(T) \leq 1 + \sqrt{\ln n(n-1)}e^{\rho^{-1}(1-\frac{1}{\ln n})T} \leq 1 + \sqrt{\ln n} n^{1-\alpha/4} = o(n)$. □

Application to particular contagion model: proofs of Corollary 12, Proposition 15, Proposition 17 and Corollary 13

Proof of Corollary 12. Taking $\rho(s) = \frac{\lambda}{\lambda+s}\rho_\alpha$, Corollary 5 rewrites

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \min_{s\geq 0} \left(\sqrt{\frac{\lambda}{s+\lambda(1-\rho_\alpha)}} e^{sT} \right). \tag{B.43}$$

The function $f(s) = \sqrt{\frac{\lambda}{s+\lambda(1-\rho_\alpha)}} e^{sT}$ admits a unique minimum in $s_{min} = \frac{1}{2T} + \lambda(\rho_\alpha - 1)$. The minimum for $s \geq 0$ is therefore met for $s = s_{min}$ if $\lambda T(1-\rho_\alpha) < \frac{1}{2}$ and $s = 0$ otherwise. The results follow immediately. □

Proof of Proposition 15. In order to prove Proposition 15, it is sufficient to show that Lemma 15 still holds for the SIR model, with $p_{ij}(t) = \beta \exp(-(\delta+\beta)t)$ for $(i,j) \in \mathscr{E}$. For $i \in \mathscr{V}$, let θ_i be the random removal time of node i. Infection times τ_i are then given by the following expression, where $\mathscr{Q}_i$ is the collection of directed paths in G from the influencers A to node i:

$$\tau_i = \min_{q\in\mathscr{Q}_i} \sum_{(j,l)\in q} \tau_{jl} \mathbb{1}_{\{\tau_{jl}<\theta_j\}} \tag{B.44}$$

Therefore $\forall i \notin A$ and $t > 0$, the functions $f_{ij}(\tau,\theta) = \{1 - \mathbb{1}_{\{\tau_j+\tau_{ji}<t\}}\mathbb{1}_{\{\tau_{ji}<\theta_j\}}\}_{j\in\mathscr{V}}$ are increasing with respect to the partial order on $\mathbb{R}^{\mathscr{E}} \times \mathbb{R}^{\mathscr{V}}$ defined for any $X^1 = (\tau_1^1, ...\tau_m^1, \theta_1^1 ...\theta_n^1) \in \mathbb{R}^{\mathscr{E}} \times \mathbb{R}^{\mathscr{V}}$ and $X^2 = (\tau_1^2, ...\tau_m^2, \theta_1^2 ...\theta_n^2) \in \mathbb{R}^{\mathscr{E}} \times \mathbb{R}^{\mathscr{V}}$ by:

$$X^1 \geq X^2 \iff \begin{cases} \tau_{ij}^1 \geq \tau_{ij}^2 & \text{for any } (i,j) \in \mathscr{E} \\ \theta_i^1 \leq \theta_j^2 & \text{for any } i \in \mathscr{V} \end{cases}. \tag{B.45}$$

Variables $(\tau_{ij})_{(i,j)\in\mathscr{E}}$ and $(\theta_i)_{i\in\mathscr{V}}$ being independent, we can still apply FKG inequality (Lemma 10) and deduce the positive correlation, for any $i \notin A$ and $t > 0$, of the random variables $\{1 - \mathbb{1}_{\{\tau_j+\tau_{ji}<t\}}\mathbb{1}_{\{\tau_{ji}<\theta_j\}}\}_{j\in\mathscr{V}}$. We then introduce, for any $(i,j) \in \mathscr{E}$:

$$\overline{\tau_{ji}} = \begin{cases} \tau_{ji} & \text{if } \tau_{ji} < \theta_j \\ +\infty & \text{if } \tau_{ji} \geq \theta_j \end{cases}. \tag{B.46}$$

It is straightforward that each $\overline{\tau_{ji}}$ is a random variable over $\mathbb{R}_+ \cup \{+\infty\}$ with probability distribution p_{ij}, and that $\overline{\tau_{ji}}$ is independent of τ_j. We also have, for any $i \notin A$, $t > 0$ and $(i,j) \in \mathscr{E}$:

$$\{1 - \mathbb{1}_{\{\tau_j + \tau_{ji} < t\}} \mathbb{1}_{\{\tau_{ji} < \theta_j\}}\} = \{1 - \mathbb{1}_{\{\tau_j + \overline{\tau_{ji}} < t\}}\} \tag{B.47}$$

Lemma 15 for the SIR case (and therefore Proposition 15 and its subsequent corollaries) are then proved from following the same steps than in the independent transmission events case, except replacing $(\tau_{ji})_{(i,j)\in\mathscr{E}}$ by $(\overline{\tau_{ji}})_{(i,j)\in\mathscr{E}}$ □

Proof of Proposition 17. $\rho(s) = \frac{\beta T_n}{1-e^{-\beta T_n}} \frac{\beta}{\beta+s}(1 - e^{-(\beta+s)T_n})\rho(\mathscr{A}) \leq \frac{\beta^2 T_n \rho(\mathscr{A})}{(1-e^{-\beta T_n})s}$, which implies $\rho^{-1}(1)T_n \leq \frac{(\beta T_n)^2 \rho(\mathscr{A})}{1-e^{-\beta T_n}}$. Let $f(x) = \frac{x^2}{1-e^{-x}}$, f is increasing and $\forall a \geq 0$, $f(x) = a \implies x \geq \sqrt{a}(1-e^{-\sqrt{a}})$. Hence, if $\limsup_{n\to+\infty} \frac{\beta T_n}{\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}(1-e^{-\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}})} < 1$, then $\exists \alpha > 0$ s.t. $\beta T_n \leq (1-\alpha)\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}(1 - e^{-\sqrt{\frac{\ln n}{2\rho(\mathscr{A}_n)}}})$, and the concavity of $1-e^{-x}$ implies that $\beta T_n \leq \sqrt{\frac{(1-\alpha)\ln n}{2\rho(\mathscr{A}_n)}}(1 - e^{-\sqrt{\frac{(1-\alpha)\ln n}{2\rho(\mathscr{A}_n)}}})$. Finally, $f(\beta T_n) \leq \frac{(1-\alpha)\ln n}{2\rho(\mathscr{A}_n)}$ and $\frac{2\rho^{-1}(1)T_n}{\ln n} \leq 1 - \alpha$. Applying Corollary 11 proves the desired result. □

Proof of Corollary 13. Taking $\rho(s) = \rho_\alpha e^{-sT_0}$, Corollary 5 rewrites

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \min_{s\geq 0} \left(\sqrt{\frac{\rho_\alpha e^{-sT_0}}{1 - \rho_\alpha e^{-sT_0}}} e^{sT} \right). \tag{B.48}$$

and $s = \frac{1}{T_0}\left(\ln \rho_\alpha - \ln(1 - \frac{T_0}{2T})\right)$ gives

$$\sigma_A(T) \leq n_0 + \sqrt{n_0(n-n_0)} \sqrt{\frac{2T}{T_0} - 1} \left(\frac{\rho_\alpha}{1 - \frac{T_0}{2T}} \right)^{\frac{T}{T_0}}. \tag{B.49}$$

The final result follows by upper bounding $\left(1 - \frac{T_0}{2T}\right)^{\frac{1}{2} - \frac{T}{T_0}}$ by $\sqrt{e}$ due to the monotonic increase of $x \to (x-1)\ln(1-\frac{1}{x})$ on $[1, +\infty[$ and its limit when $x \to +\infty$. □

References

[1] Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

[2] Syndicat des régies internet. Qunizième observatoire de l'epub sri. http://www.sri-france.org/etudes-et-chiffre-cles/observatoire-de-le-pub-sri/15eme-observatoire-de-pub-sri/. Accessed: 2016-03-31.

[3] Robert B Settle and Linda L Golden. Attribution theory and advertiser credibility. *Journal of Marketing Research*, pages 181–185, 1974.

[4] Emmanuel Bacry, Khalil Dayri, and Jean-Francois Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12, 2012.

[5] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[6] Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

[7] Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1301–1309, 2013.

[8] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[9] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

[10] Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*, 2014.

[11] Justin Kirby and Paul Marsden. *Connected marketing: the viral, buzz and word of mouth revolution*. Elsevier, 2006.

[12] Michael Trusov, Randolph E Bucklin, and Koen Pauwels. Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of marketing*, 73(5):90–102, 2009.

[13] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[14] David Kempe, Jon Kleinberg, and Éva 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*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[15] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009.

[16] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010.

[17] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011.

[18] Kouzou Ohara, Kazumi Saito, Masahiro Kimura, and Hiroshi Motoda. Predictive simulation framework of stochastic diffusion model for identifying top-k influential nodes. In *Asian Conference on Machine Learning*, pages 149–164, 2013.

[19] Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

[20] Seth A. Myers and Jure Leskovec. On the convexity of latent social network inference. In *Advances In Neural Information Processing Systems*, pages 1741–1749, 2010.

[21] P Erdős and A Rényi. On the evolution of random graphs. *Selected Papers of Alfréd Rényi*, 2:482–525, 1976.

[22] Béla Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*, 286(1):257–274, 1984.

[23] Tomasz Łuczak. Component behavior near the critical point of the random graph process. *Random Structures & Algorithms*, 1(3):287–310, 1990.

[24] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[25] Béla Bollobás, Christian Borgs, Jennifer Chayes, Oliver Riordan, et al. Percolation on dense graph sequences. *The Annals of Probability*, 38(1):150–183, 2010.

[26] William O Kermack and Anderson G McKendrick. Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

[27] Moez Draief, Ayalvadi Ganesh, and Laurent Massoulié. Thresholds for virus spread on networks. *Annals of Applied Probability*, 18(2):359–378, 2008.

[28] Kenrad E Nelson. Epidemiology of infectious disease: general principles. *Infectious Disease Epidemiology Theory and Practice. Gaithersburg, MD: Aspen Publishers*, pages 17–48, 2007.

[29] Manuel Gomez-Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 561–568, 2011.

[30] Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. Uncover topic-sensitive information diffusion networks. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 229–237, 2013.

[31] Manuel G Rodriguez and Bernhard Schölkopf. Influence maximization in continuous time diffusion networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 313–320, 2012.

[32] Nan Du, Le Song, Manuel Gomez-Rodriguez, and Hongyuan Zha. Scalable influence estimation in continuous-time diffusion networks. In *Advances in Neural Information Processing Systems*, pages 3147–3155, 2013.

[33] Remi Lemonnier and Nicolas Vayatis. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes. In *Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, 2014.

[34] Rémi Lemonnier, Kevin Scaman, and Argyris Kalogeratos. Multivariate hawkes processes for large-scale inference. *arXiv preprint arXiv:1602.08418*, 2016.

[35] David Oakes. The Markovian self-exciting process. *Journal of Applied Probability*, pages 69–77, 1975.

[36] Thomas Josef Liniger. *Multivariate Hawkes processes*. PhD thesis, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403, 2009.

[37] David Vere-Jones. Earthquake prediction – statistician's view. *Journal of Physics of the Earth*, 26(2):129–146, 1978.

[38] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[39] Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont, and Christine Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *Journal of Mathematical Neurosciences*, page 4:3, 2014.

[40] Eymen Errais, Kay Giesecke, and Lisa R Goldberg. Pricing credit from the top down with affine point processes. *Numerical Methods for Finance*, pages 195–201, 2007.

[41] Luc Bauwens and Nikolaus Hautsch. *Modelling financial high frequency data using point processes*. Springer, 2009.

[42] Emmanuel Bacry, Sylvain Delattre, Marc Hoffmann, and Jean-François Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1):65–77, 2013.

[43] Aurélien Alfonsi and Pierre Blanc. Dynamic optimal execution in a mixed-market-impact hawkes price model. *Finance and Stochastics*, 20(1):183–218, 2015.

[44] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[45] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[46] Yoshiko Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

[47] Erik Lewis and George Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Joint Statisticals Meetings 2011*, 2011.

[48] Erik Lewis, George Mohler, P Jeffrey Brantingham, and Andrea L Bertozzi. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25(3):244–264, 2011.

[49] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes*. Springer, 2007.

[50] Per Kragh Andersen. *Statistical models based on counting processes*. Springer, 1993.

[51] Elliott Ward Cheney and Elliott Ward Cheney. *Introduction to approximation theory*, volume 3. McGraw-Hill New York, 1966.

[52] Stephen Poythress Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[53] MJD Powell. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. *Nonlinear programming*, 9:53–72, 1976.

[54] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[55] Yurii Nesterov, Arkadii Semenovich Nemirovskii, and Yinyu Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[56] D Seung and L Lee. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

[57] Thibault Jaisson. *Market activity and price impact throughout time scales*. PhD thesis, Ecole Polytechnique, 2015.

[58] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[59] Otto Szász. Über die approximation stetiger funktionen durch lineare aggregate von potenzen. *Mathematische Annalen*, 77(4):482–496, 1916.

[60] Dilcia Pérez and Yamilet Quintana. A survey on the weierstrass approximation theorem. *Divulgaciones Matemáticas*, 16(1):231–247, 2008.

[61] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, 2014.

[62] Remi Lemonnier, Kevin Scaman, and Nicolas Vayatis. Tight bounds for influence in diffusion networks and application to bond percolation and epidemiology. In *Advances in Neural Information Processing Systems*, pages 846–854, 2014.

[63] Kevin Scaman, Rémi Lemonnier, and Nicolas Vayatis. Anytime influence bounds and the explosive behavior of continuous-time diffusion networks. In *Advances in Neural Information Processing Systems*, pages 2026–2034. 2015.

[64] Rémi Lemonnier, Kevin Scaman, and Nicolas Vayatis. Spectral bounds in random graphs applied to spreading phenomena and percolation. *arXiv preprint arXiv:1603.07970*, 2016.

[65] Mark Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[66] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

[67] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[68] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145, 2002.

[69] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[70] Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics probability and computing*, 7(3):295–305, 1998.

[71] Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press Oxford, 2003.

[72] Mehrdad Farajtabar, Nan Du, Manuel Gomez-Rodriguez, Isabel Valera, Hongyuan Zha, and Le Song. Shaping social activity by incentivizing users. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2014.

[73] Manuel Gomez-Rodriguez, Le Song, Hadi Daneshmand, and B. Schoelkopf. Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm. *Journal of Machine Learning Research*, 2015.

[74] Jean Pouget-Abadie and Thibaut Horel. Inferring graphs from cascades: A sparse recovery framework. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 977–986, 2015.

[75] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 420–429. ACM, 2007.

[76] Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[77] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 2. Siam, 2000.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Applications des processus stochastiques aux enchères en temps réel et à la propagation d'information dans les réseaux sociaux

Mots clefs : Processus de Hawkes, Real-time bidding, Processus diffusifs sur les graphes, Cascades d'information, Maximisation d'influence, Marketing viral, Bond Percolation, Modèle SIR, Inférence non-paramétrique à grande échelle

Résumé : Dans cette thèse, nous étudions deux applications des processus stochastiques au marketing internet. Le premier chapitre s'intéresse au scoring d'internautes pour les enchères en temps réel. Ce problème consiste à trouver la probabilité qu'un internaute donné réalise une action d'intérêt, appelée conversion, dans les quelques jours suivant l'affichage d'une bannière publicitaire. Nous montrons que les processus de Hawkes [1] constituent une modélisation naturelle de ce phénomène mais que les algorithmes de l'état de l'art ne sont pas applicables à la taille des données typiquement à l'œuvre dans des applications industrielles. Nous développons donc deux nouveaux algorithmes d'inférence non-paramétrique qui sont plusieurs ordres de grandeurs plus rapides que les méthodes précédentes. Nous montrons empiriquement que le premier a de meilleures performances que les compétiteurs de l'état de l'art, et que le second permet une application à des jeux de données encore plus importants sans payer un prix trop important en terme de pouvoir de prédiction. Les algorithmes qui en découlent ont été implémentés avec de très bonnes performances depuis plusieurs années à 1000mercis, l'entreprise spécialiste du marketing interactif étant le partenaire industriel de cette thèse CIFRE, où ils ont beaucoup apporté en production.

Le deuxième chapitre s'intéresse aux processus diffusifs sur les graphes qui constituent un outil important pour modéliser la propagation d'une opération de marketing viral sur les réseaux sociaux. Nous établissons les premières bornes théoriques sur le nombre total de nœuds atteint par une contagion dans le cadre de graphes et dynamiques de diffusion quelconques, et montrons l'existence de deux régimes bien distincts : le régime sous-critique où au maximum $O(\sqrt{n})$ nœuds seront infectés, où n est la taille du réseau, et le régime sur-critique ou $O(n)$ nœuds peuvent être infectés. Nous étudions également le comportement par rapport au temps d'observation T et mettons en lumière l'existence de temps critiques en-dessous desquels une diffusion, même sur-critique sur le long terme, se comporte de manière sous-critique. Enfin, nous étendons nos travaux à la percolation et l'épidémiologie, où nous améliorons les résultats existants.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Title : Application of stochastic processes to real-time bidding and diffusion processes on networks

Keywords : Multivariate Hawkes processes, Real-time bidding, Diffusion processes on graphs, Information cascades, Influence maximization, Viral marketing, Bond percolation, SIR Model, Large-scale nonparametric inference

Abstract : In this thesis, which is the result of a CIFRE collaboration with the pioneering marketing agency 1000mercis, we study two applications of stochastic processes in internet marketing. The first chapter focuses on internet user scoring for real-time bidding. This problem consists in finding the probability for a given user to perform an action of interest, called conversion, in the next few days. We show that Hawkes processes [1] are well suited for modeling such phenomena but that state-of-the-art algorithms are not applicable to the size of data sets typically involved in industrial applications. We therefore develop two new algorithms able to perform nonparametric multivariate Hawkes process inference orders of magnitude faster than previous methods. We show empirically that the first one outperforms state-of-the-art competitors, and the second one scales to very large datasets while maintaining very high prediction power. The resulting algorithms have been integrated in production and are used on everyday basis with remarkable performance in 1000mercis, where they became an important business asset. The second chapter focuses on diffusion processes on graphs, an important tool for modelizing the spread of a viral marketing operation over social networks. We derive the first theoretical bounds for the total number of nodes reached by a contagion for general graphs and diffusion dynamics. For any graph of size n, we show the existence of two distinct regimes: the sub-critical one where at most $O(\sqrt{n})$ nodes are infected, and the super-critical one where $O(n)$ nodes can be infected. We also study the behavior w.r.t. to the observation time T and reveal the existence of critical times under which a long-term super-critical diffusion process behaves sub-critically. Finally, we extend our work to different application fields, and improve state-of-the-art results in percolation and epidemiology.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France