Predicting and containing epidemic risk using friendship networks

Lorenzo Coviello, Massimo Franceschetti, Manuel García-Herranz, Iyad Rahwan

Abstract—Physical encounter is the most common vehicle for the spread of infectious diseases, but detailed information about said encounters is often unavailable because expensive, unpractical to collect or privacy sensitive. The present work asks whether the friendship ties between the individuals in a social network can be used to successfully predict and contain epidemic risk. Using a dataset from a popular online review service, we build a time-varying network that is a proxy of physical encounter between users and a static network based on their reported friendship – the encounter network and the friendship network. Through computer simulation, we compare infection processes on the resulting networks and show that friendship provides a poor identification of the individuals at risk if the infection is driven by physical encounter. This result is not driven by the static nature of the friendship network opposed to the time-varying nature of the encounter network, as a static version of the encounter network provides more accurate prediction of risk than the friendship network. Despite this limit, the information enclosed in the friendship network can be leveraged for monitoring and containment of epidemics. In particular, we show that periodical and relatively infrequent monitoring of the infection on the encounter network allows to correct the predicted infection on the friendship network and to achieve satisfactory prediction accuracy. In addition, the friendship network contains valuable information to effectively contain epidemic outbreaks when a limited budget is available for immunization.

I. INTRODUCTION

The forecast and mitigation of epidemics is a central theme in public health [17], [25], [26], [27], [34], [42], [56]. Events such as the recent ebola epidemic constantly drive the attention and resources of governments, institutions such as the World Health Organization, and the research community [31], [35], [51], [55], [58], [71]. The study of infectious processes on real-world networks is of interests to diverse disciplines, and similar models have been proposed to characterize the spread of information, behaviors, cultural norms, innovation, as well as the diffusion of computer viruses [30], [54], [69], [72], [84]. Therefore, epidemiologists, computer scientists and social scientist have joint forces in the study of these phenomena. Due to the impossibility to study the spread of infectious diseases through controlled experiments, modeling efforts have prevailed [36], [47], [50], [68], [69]. Recently, advancements in computation tools determined the emergence of data-driven simulations in the study of dynamical processes [86].

The spread of an infection over a real-world network is determined by the interplay of two processes: the dynamics of the network, whose edges change over time according to the encounters between individuals; and the dynamics on the network, whose nodes can infect each other after an encounter. When the two dynamics operate at comparable time scales, the time-varying nature of the network cannot be ignored [33], [41], [48], [70], [76] and specifically devised control strategies are necessary [53]. Aggregating the dynamics of the edges into a static version of the network can provide useful insights [21] but can introduce bias [39], [70]. Empirical work suggests that bursty activity patterns slow down spreading [46], [80], [85], but temporal correlations seem to accelerate the early phase of an epidemic [45], [74].

Physical encounter is the most common vehicle for the spread of infectious diseases (as in the case of airborne diseases), and detailed information about said encounters is fundamental to monitor and contain outbreaks. Various sources of data can serve as a proxy of physical encounter – checkins on social networking platforms [14], [63], [64], traffic records [5], [82], [83], phone call records [32], [38], [65], wearable sensor data [12], [37], [43], [66], [75], [80], geographical and non-geographical information shared online [4], [13], surveys and diaries of daily contact [20], [59], [60], and multiplex data [81]. However, pervasive and detailed information is rarely available and might be expensive and unpractical to collect (as in the case of sensor technologies [12], [75], [80]), prone to errors (as in the case of survey data [19], [73]), and privacy-sensitive [2], [7], [8], [18], [49], [52], [77], [88]. Researchers have to rely on the information in their possession, and in this paper we consider friendship ties between the users of an online social network. Location data from cell phone records and online social networks has shown that social relationships can partially explain the patterns of human mobility [15]. Contact tracing based on phone communication activity has been proposed as a method to reduce the final size of epidemic outbreaks [23]. Networks generated from wearable sensor measurements, diaries of daily contacts, online links and self-reported friendship present similar structural properties [57], but contacts of short duration recorded by wearable sensors might not be reported in surveys [78]. In general, it is not clear whether and within which limits friendship can be considered a reliable proxy of physical encounter, as a process spreading from an initial seed, or “patient zero”, can only reach the nodes within its set of influence through paths that respect time ordering [40].

Given an infection spreading over a social network via physical encounter, the present work asks whether the friendship ties between the same individuals successfully predict who is at risk, and how this information can be leveraged to monitor and contain epidemic outbreaks. Using the Yelp
Dataset Challenge dataset, we build a time-varying network that is a proxy of physical encounter between users and a static network based on their reported friendship – the encounter network and the friendship network. For comparison, we also consider a static version of the encounter network, in which temporal information is ignored. Through computer simulation, we compare the evolution of Susceptible-Infected (SI) processes [3] on the different networks, in terms of the sets of infected individuals. First, given epidemic processes spreading independently on the two networks but initiated at the same seed, we analyze infections from a microscopic point of view. Despite distance on the friendship network is correlated to risk of infection on the encounter network, the set of nodes infected on the friendship network is not a good approximation of those infected in the encounter network. Then, we consider how this limit can be overcome. On the one hand, we show that periodical yet relatively infrequent observation of the real infection allows to reach and maintain high prediction accuracy using the friendship network. On the other hand, we show that friendship ties allow to effectively allocate a limited immunization budget, in order to reduce the risk of an outbreak.

II. DATASET

The Yelp Dataset Challenge dataset (Round 5) consists in 1,569,264 reviews and 495,107 tips to 61,184 businesses (in 10 cities around the world) posted by 366,715 users over more than 10 years. Within this period, we consider 1,469 consecutive days ranging from 1/1/2011 to 1/8/2015 (period of observation \( T \)), as reviews and tips before year 2011 are less numerous. Each review and tip includes the user who posted it, the reviewed business, and the date it was posted. Yelp users can form friendship ties between each other, and the list of friends of each user is included in the dataset. The encounter network (SI Section 2 for details) is a sequence of networks \( \{ N_E(t), t \in T \} \), where \( N_E(t) \) is the undirected network of encounters on day \( t \). We assume that two users \( x, y \), encountered on day \( t \), and \( (x, y) \in N_E(t) \), if they posted reviews or tips to the same business on the same day (133,038 users had at least one encounter during \( T \)). For node \( s \), let \( t_0(s) \) be the smallest \( t \) such that \( s \) is connected in \( N_E(t) \) (first encounter). The static version of the encounter network \( N_S \) ignores the temporal information and \( (x, y) \in N_S \) if \( (x, y) \in N_E(t) \) for some \( t \) (we consider the 113,187 nodes in the largest connected component). The friendship network \( N_F \) is the undirected, static network of friendship ties (168,923 nodes in the largest connected component).

III. INFECTION DYNAMICS

We consider Susceptible-Infected processes [3], in which nodes do not recover (SI Section 3 for details). Given a temporal network \( \{ N(t), t \in T \} \), let \( I(t) \) be the set of infected nodes at time \( t \). Each simulation is initiated at a single seed \( I(0) = \{ s \} \), selected uniformly at random from the nodes in \( N_S \cap N_F \) such that \( t_0(s) \leq 500 \), unless differently stated (to allow the infection to start early enough on the encounter network). At time \( t \), each node \( x \) not in \( I(t-1) \) becomes infected with probability \( \max\{ \beta d_x(t), 1 \} \), where \( 0 \leq \beta \leq 1 \) is the infection rate and \( d_x(t) \) is the number of neighbors of \( x \) in \( N(t) \) that are in \( I(t-1) \). For the encounter network, \( N(t) = N_E(t) \) and the infection spreads from time \( t_0(s) \) until \( \max T \). For the static (resp. friendship) network \( N(t) = N_S \) (resp. \( N_F \)) for all \( t \), and the infection spreads from time zero until all nodes are infected. We consider different values of \( \beta \) for the different networks, due to their different connectivity (usually \( \beta = 0.5 \) on the encounter network, and \( \beta = 0.01 \) on the other networks).

IV. RESULTS

Throughout, we assume that real infections spread on the encounter network. Infection spreading on the friendship network and the static version of the encounter network are predicted infections. Given an infection spreading on the encounter network, a node’s risk of becoming infected is correlated to its distance on the friendship network from the infected seed. Figure 1 plots the probability of nodes’ infection as a function of this distance, for different values of the infection rate \( \beta \) (10,000 simulations for each \( \beta \)). In order to evaluate how accurately the friendship network predicts epidemic risk at a microscopic level, we consider infection processes initiated at seeds present in both networks and spreading independently of each other, and compare the sets of infected nodes (SI Section 5 for details). The unpredictability of epidemic risk is due to the structural properties of the different networks as well as to the randomness of the infection processes. Therefore, for each of 5,000 seeds, we consider two independent infection processes on the encounter network, one on the friendship network and one on the static version of the encounter network (indexed by \( E_1, E_2, F \) and \( S \), respectively). For target size \( m \) and infection \( A \) initiated at seed \( s \), let \( I_A(m; s) \) be the first time at which at least \( m \) nodes are infected (the quantity might be undefined). Let \( I_A(m; s) \) be the corresponding infected set, whose size is at least \( m \) when defined. Consider two infections \( A \) and \( B \) initiated at the same seed \( s \) and possibly spreading on two different networks. If \( I_A(m; s) \) and \( I_B(m; s) \) are defined, their Jaccard similarity is given by

\[
J_{A,B}(m; s) = \frac{|I_A(m; s) \cap I_B(m; s)|}{|I_A(m; s) \cup I_B(m; s)|}.
\]

These measures allow to characterize how accurately the friendship network and the static network predict epidemic risk on the encounter network (see SI for details and analyses involving different metrics). For each of 5000 selections of a random single seed, two simulations on the encounter network, one on the static network and one on the friendship network are run independently. Results are shown in Figure 2 and SI Section 5. The left panel shows the metrics \( J_{E_1,E_2}(m; s) \), the baseline unpredictability due solely to the randomness of independent processes initiated at the same seed and spreading.
Therefore, we also consider a rescaled version of $J$ (even between independent processes initiated at the same seed identification of the individuals at risk if the infection is driven at the same seed level (SI Section 10 and 11), friendship provides a poor network. Our analyses show that, despite friendship differences between the networks affect the unpredictability how the randomness of the infection process and the structural networks produce similar epidemic dynamics at the macro level (SI Section 10 and 11). The analyses reported thus far might erroneously lead to negative conclusions about the possibility of using the friendship network for prediction and containment of epidemic risk. From the point of view of outbreak containment, we consider a scenario in which a fixed budget is available for immunization (e.g., limited amount of vaccine) and must be effectively allocated in order to contain an epidemic on the encounter network. In contrast to purely random immunization, we consider a strategy that selects random friends of randomly chosen individuals for immunization (friend immunization). Name-a-friend methods have already been proposed to predict the peak of an epidemic outbreak [16] and the spread of information online [28], and are motivated by the “friendship paradox” – the average friend of an individual is more connected than the average individual [24]. The method results in a more effective use of the immunization budget, substantially increasing the probability that an infection dies out in its early stage (Figure 4) and strongly reducing the final infection size (Figures 5 and S18-S20) with respect to random immunization. Moreover, it only requires a small additional cost to obtain the same effect as an ideal strategy that targets future encounters rather than friends (encounter immunization, see SI Section 9).
Fig. 2. Predictability of nodes’ epidemic risk. For each of 5000 selections of a random single seed, two simulations on the encounter network, one on the static network and one on the friendship network are run independently. The similarities $J_c(m,s)$ of the infected sets are shown for different pairs of networks and different target infection size $m$. On the x-axis, observations for a given value of $m$ form a block with a constant color (within the block, the x position is irrelevant). Black points represent the averages of the metrics over all observations such that the metrics are defined, and bars represent standard deviations.

Fig. 3. Periodical correction of risk prediction using the friendship network. Shown here is the Jaccard similarity between the predicted infected set on the friendship network and the real infected set on the encounter network before each correction, for different values of the observation window $W$. For each $W \in \{10, 20, 50\}$, 6000 single seeds are selected at random, and for each seed one simulation on the encounter network and one (with correction) on the friendship network are run.

Fig. 4. Fraction of infections that do not die out in the early stage as a function of immunization budget $b$ and immunization method. For each immunization type and $b \in \{1\%, 2\%, 5\%, 10\%, 15\\}$, 5000 simulations on the encounter network initiated at random single seeds are run. The x-axis shows $b$, the y-axis shows the fraction of infections whose final size is above 0.1% of the entire population (taken as an indicator that the infection did not die out). Immunization budget is expressed as a fraction $b$ of the entire population. For $b \in \{1\%, 2\%, 5\%, 10\%, 15\\}$, Figure 4 shows the fraction of infections above 0.1% of the entire population as a function of $b$ for all considered immunization methods (5000 simulations for each immunization method and $b$). The 0.1% target infection size is an indicator that the infection did not die out. The trend in Figure 4 is captured by a linear model with interactions between immunization type and immunization budget ($R^2 = 0.98$). Each 1% increase of the immunization budget determines: a 0.5% decrease in the fraction of infections above the 0.1% target for random immunization (p-value = 0.0299); an additional 2.36% decrease for friend immunization (p-value = 2.77 \cdot 10^6); and an additional 3.5% decrease for encounter immunization (p-value = 4.03 \cdot 10^8). Regarding the size of the infected population, Figure 5 shows (for $b = 5\%$) the fraction of infections with final rate above given targets (5%, 10%, 15% of the entire population in the three panels) as a function of the infection start time $t_0(s)$ for all immunization methods. Friend immunization provides a large advantage with respect
to random immunization and, despite its simplicity, requires a relatively small additional cost to reach the same effectiveness as the benchmark of encounter immunization.

V. Discussion

Since seminal work on the structure and growth of complex networks [6], [22], [87], interdisciplinary research has shown that biological networks, social networks and the Internet are governed by similar rules [1], [9], [44], [61] and share similar structure [29], [62], [67]. Very similar models have been proposed to characterize the spread of epidemics, information, behaviors, and cultural norms. Despite the macroscopic similarity between processes spreading on different networks, our work shows that the differences in local connectivity determined by the two definitions of edges result in striking differences between the dynamics at a microscopic level, that prevent the identification of the nodes at risk using a friendship network. However, this limitation can be overcome by periodical yet infrequent monitoring of the real infection on the encounter network. In addition, in the context of immunization with limited budget, simply asking individuals to name a friend enables the effective use of the available resources, increasing the probability that the infection dies out in its early stage and reducing the final size of the infected population.

We considered reviews as a proxy of physical encounter – an edge is active between two users on day t if they posted a review to the same business on day t. This constitutes an approximation to real physical encounter, that would requires users to visit (rather than review) a business at about the same time. This approximation is justified as the time of a review is a proxy of the time of the visit to a business, and infections do not necessarily require direct physical contact. For example, in the case of airborne transmission, particles can remain suspended in the air for hours after an infected individual has occupied a room [10]. In the context of our dataset, after an infected user visits a business, the infection might spread to customers who visit the business later on the day. Also, the virus can infect customers which are not included in the dataset, and from them can infect another user who visits the business in a later moment.

When it is known who is infected or likely to become infected (e.g., individuals traveling to certain countries who might have come in contact with a pathogen), accurate prediction of the individuals at risk of contagion would allow targeted monitoring and immunization. Despite friendship and other social relationships might be informative about the encounters between individuals, our work suggests that they do not always give a complete picture of the paths a pathogen might take. Information to predict future encounters between individuals is likely to be unavailable, at least at a detailed level. However, a feasible approach could use past encounter as a proxy of future encounter. In fact, it is known that human mobility and encounter present high spatial and temporal regularity and predictability [11], [32], [79], [83]. From a practical perspective, networks based on social relationships (such as a friendship network) might be complemented by information about past encounter. Our simulations are based on a large dataset that allowed us to build a static friendship network and a time-varying encounter network that is a candidate vehicle for the spread of a pathogen. The dataset considers a large number of individuals and spans several years of activity. In general, other datasets might be available and allow similar analyses. Friendship networks whose edges have a different semantic than that considered in the present work might lead to different observations.

References

For each immunization type, 5000 simulations on the encounter network are initiated at random single seeds. Each panel considers a target infection size for which the final size is above the given target.

Fig. 5. Final infection size as a function of the immunization method and the infection start time. Given immunization budget $b$ -axis shows the infection start time $t_0(s)$ of seed $s$, the y-axis shows the fraction of infections whose final size is above the given target.


