Predicting and containing epidemic risk using on-line friendship networks

Lorenzo Coviello
University of California San Diego
La Jolla, CA
lorenzocoviello@gmail.com

Massimo Franceschetti
University of California San Diego
La Jolla, CA
lorenzocoviello@gmail.com

Manuel Garcia Herranz
Unicef Innovation Unit
New York, NY

Iyad Rahwan
MIT Media Lab
Cambridge, MA

ABSTRACT

To what extent can online social ties predict who is at risk of infection? Physical encounter between humans is a common vehicle for the spread of airborne diseases, but collecting detailed information can be expensive or might invade privacy. The present work asks whether the online social ties reported by the individuals in a social network can be used to predict and contain epidemic risk. Using a dataset from a popular online review service which includes over 100k users and spans 4y of activity, we build a time-varying network that is a proxy of physical encounter between its users and a static network based on their reported online friendship – the encounter network and the friendship network. We compare stochastic infection processes on the two networks through computer simulation, considering infections on the encounter network as the benchmark. We show that the friendship network is useful for identifying the individuals at risk of infection, despite providing lower accuracy than an ideal case in which the encounters are known. The limited prediction accuracy of the friendship network is not only driven by its static nature, since a static version of the encounter network provides more accurate prediction of risk. Periodical and relatively infrequent reports of the infection spreading on the encounter network allow to correct the infection predicted by a process spreading on the friendship network, and to achieve high prediction accuracy. In addition, the friendship network contains valuable information to effectively contain epidemic outbreaks when a limited budget is available for immunization. The strategy to immunize random friends of random individuals achieves the same performance as knowing individuals' encounters at a relatively small additional cost, even if the infection spreads on the encounter network.

ACM Reference Format:
Lorenzo Coviello, Massimo Franceschetti, Manuel Garcia Herranz, and Iyad Rahwan. 2018. Predicting and containing epidemic risk using on-line friendship networks. In *Proceedings of ACM International Conference on Web Search and Data Mining, Los Angeles, CA USA, February 2018 (WSDM'18),* 9 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The forecast and containment of epidemics is a central theme in public health [11, 16, 20, 32]. Events such as the recent ebola epidemic constantly drive the attention and resources of institutions such as the World Health Organization, governments, and researchers [13, 27, 31]. Beside biological epidemics, the study of infectious processes is of broad interest as it models the spread of information, behaviors, cultural norms, innovation, as well as the diffusion of computer viruses [30, 35, 37, 43]. As it is impossible to study the spread of infectious diseases through controlled experiments, and thanks to advancements in computation, modeling efforts have prevailed [8, 23, 26].

The spread of an infection over a real-world network is determined by the interplay of two processes: the structural dynamics *of* the network, whose edges change over time, and the infection dynamics *on* the network, whose paths are constrained by the realization of the former process. When the two dynamics operate at comparable time scales, the time-varying nature of the network cannot be ignored [15, 19, 24, 36, 39], specifically devised control strategies are necessary [29], and aggregating the dynamics of the edges into a static version of the network can introduce bias [17, 36]. Empirical work suggests that bursty activity patterns slow down spreading [22, 41, 44], but temporal correlations seem to accelerate the early phase of an epidemic [21, 38].

Physical encounter between humans is a common vehicle for the spread of airborne diseases. Thus, knowledge of the patterns of human encounter is fundamental for monitoring and containing outbreaks. Various sources of data can be used as a proxy of physical encounter – check-ins on social networking platforms, traffic records, phone call records, and wearable sensor data constitute examples. However, pervasive and detailed information is rarely available and might be expensive and unpractical to collect (as in the case of sensor technologies), prone to errors (as in the case of survey data), and collection might invade privacy [7, 25, 28].

In the absence of information about human encounter, social network information obtained via mining of massive on-line social platforms might be useful to design strategies for containment of epidemic outbreaks. To what extent can online social ties predict who is at risk of infection? How can information about such relationships help monitor and contain epidemic outbreaks? Recent work has suggested that communication traces obtained from mobile phones might help reducing the expected size of an epidemic [9]. In addition, networks generated from wearable sensor measurements, diaries of daily contacts, online links and self-reported friendship present similar structural properties [33]. However, it is unclear whether these global structural properties are representative of similarities

WSDM'18, February 2018, Los Angeles, CA USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

in epidemic processes at the microscopic scale. In general, it is not even clear whether 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 [18].

In this work, we consider the prediction of epidemic risk at the individual level. In particular, we use the friendship ties between the individuals in a social network to predict who has a high probability of becoming infected, given an infection driven by physical encounter initiated at a known infection seed. The ability of identifying who is at risk of infection is critical to inform containment procedures, such as the immunization of particular groups, or the mobilization of treatment facilities to specific communities at risk.

Using a dataset from the popular online review service Yelp (We consider the Yelp Dataset Challenge dataset, Round 5: www.yelp.com/dataset_challenge) which includes over 100k users and spans 4y of activity, 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 study Susceptible-Infected processes [2] spreading on the different networks and compare the sets of infected individuals, assuming that real infections spread on the encounter network and that the friendship network is available and used for prediction purposes. Considering simulated infections on the time-varying encounter network as the benchmark, we quantify the extent to which the friendship network and the static version of the encounter network provide good enough prediction.

Given epidemic processes spreading independently on the encounter and friendship networks but initiated at the same seed, we show that the friendship network contains useful information for predicting epidemic risk at the individual level. In particular, the set of nodes infected by processes spreading on the friendship network approximates those infected by processes spreading on the encounter network substantially better than random guessing. In addition, given epidemics spreading on the encounter network, we report a correlation between a node's probability of becoming infected and its distance from the infection seed on the friendship network. These are important results, as in practice it might be feasible to track friendship or other forms of static relationship, while infeasible to track or predict physical encounter.

However, the prediction accuracy obtained with the friendship network does not come close to an ideal case in which the encounters are knows (usually not the case in practice). Even if the stochasticity of the infection process certainly contributes to the unpredictability of risk at the individual level, the difference between the two networks plays a major role. In fact, two independent infections spreading on the encounter network and started at the same seed have on average substantially higher similarity than two infections spreading on the two different networks. Moreover, this result is not only driven by the static nature of the friendship network as opposed to the time-varying nature of the encounter network, since a static version of the encounter network provides more accurate prediction of risk than the friendship network.

From a practical point of view, reported friendship ties can inform monitoring and containment of epidemic outbreaks. On the one hand, we show that periodical yet relatively infrequent observation of the benchmark infection boosts the accuracy of risk prediction using the friendship network. In particular, we consider a scenario in which the encounter network is still unknown, but the set of infected nodes is observed periodically. In the case of real epidemics, reports of the infected population are usually available at regular intervals, daily, weekly or monthly, in the form of situational reports or through case management systems. After each observation, the set of infected individuals estimated by running the process on the friendship network is updated to match the set of individuals infected by a process spreading on the encounter network. By comparing the predicted infected set (obtained with the friendship network and periodical updates) and the benchmark infected set (obtained with the encounter network) immediately before each update, we show that a high level of accuracy is reached and maintained even with infrequent observations.

On the other hand, we show that online friendship ties allow to effectively allocate a limited immunization budget in order to reduce the risk of an outbreak, even if the infection spreads on the encounter network. In particular, we consider the strategy of providing immunization to random friends of randomly selected individuals, motivated by the "friendship paradox" [6, 10, 12], according to which the average individual in a network is less connected than the average friend. Compared to a basic strategy that provides immunization to randomly selected individuals, the proposed strategy increases the probability that an infection dies out in its early stages, and always reduces the size of the infected population. Its implementation only requires individuals to name a friend and avoids computing metrics such as degree and centrality. Despite its simplicity, it only requires a relatively small additional cost to provide the same effectiveness as a strategy that immunizes encounters of random individuals (which would therefore require knowledge of the encounter network).

2 DATASET

The Yelp Dataset Challenge dataset (Round 5, www.yelp.com/dataset_challenge) 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 a period spanning over than 10 years. Within this period, we consider $1,469$ consecutive days ranging from 1/1/2011 to 1/8/2015, as reviews before 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. Time information about the formation of friendship ties is not available. Using the dataset, we define two networks, called the friendship network and the encounter network respectively.

Let U be the set of users, $F \subseteq U \times U$ be the set of friendship ties, B the set of businesses, T be the set of days, $R \subseteq U \times B \times T$ be the set of reviews and tips (which we will refer to as reviews). For each user $u \in U$ let $F_u \subset U$ be the set of friends of u. Therefore $F \quad \cup_{u \in U}\{u, v : v \in F_u\}$. Each review (or tip) $r \in R$ is a triple u, b, t where $u \in U, b \in B, t \in T$.

2.1 The friendship network

Of all users, $174,100$ have at least one friend, with an average number of friends per user, or friend degree, 14.8.

Let $N_F \;\; U, F$ be the static friendship network. As we consider processes spreading between connected nodes, connectedness is the key property of the networks. Therefore, we restrict our attention to the giant component, as users outside giant components form small components whose dynamics are not relevant. The giant component defined by friendship includes 168, 923 users (whereas the second largest component has 8 users). In what follows, we will identify N_F with its giant component. Observe that this network is static, as its edges do not change over time.

2.2 The encounter network

The most common vehicle for the spread of infectious diseases is physical contact (rather than friendship) between individuals. Strictly speaking, two users in U encountered on a given day t if they visit the same business on day t at the same time. In the present work, we use reviews as a proxy of physical encounter: an edge is active between two users in U on day t if they posted a review to the same business on day t. This constitutes an approximation to real physical encounter, which 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 the element that spreads over a network (e.g., a virus or an opinion) does 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 individuals has occupied a room [4]. In the context of our dataset, after an infected user visits a business, the infection might spread to customers who visit the business later in 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.

For each $t \in T$, $Ut \;\; \{u \in U : u, b, t \in R$ for some $b \in B\}$ is the set of users who wrote a review on day t. We refer to Ut as the active users on day t.

For each $t \in T$ and $u \in Ut$, $E_u t \;\; \{v \in Ut, v \neq u : u, b, t \in R$ and $v, b, t \in R$ for some $b \in B\} \subseteq U$ is the set of encounters of user u on day t (i.e., users who visited at least one of the businesses visited by u). $Et \;\; \cup_{u \in U}\{u, v : v \in E_u t\} \subseteq U \times U$ is the set of encounters on day t.

For each $t \in T$, let $N_E t \;\; U, Et$ be the network defined by the encounters on day t. Observe that the node set in the definition is U rather than Ut. The *encounter network* is the sequence $\{N_E t\}_{t \in T}$. As connectedness is the key property in a spreading process, we consider the 133, 038 users who had at least one encounter during T.

Despite friend degree and encounter degree are correlated (Pearson product-moment correlation 0.3416, p-value $< 2.2 \cdot 10^{-16}$), the similarity of the sets of the friends and encounters of an individual is low. Figure 1 shows the cumulative distribution function of the Jaccard similarity of the set of friends and the set of encounters of all users in the dataset (left panel), of all users in the giant component of the network defined by all friendship ties (center panel), and of all users in the giant component of the network defined by all encounters (right panel). Considering the 72, 786 users with at least one friend and one encounter, the average Jaccard similarity of their encounter and friend sets is 0.01716, with only 9, 527 of them with a value different than zero. Looking at the giant component of the network defined by all friendship ties, the users with nonzero encounters have average Jaccard similarity of their encounter and friend set of 0.1306, with only 9, 022 users with a nonzero value. Looking at the giant component of the network defined by all encounters, the users with nonzero friends have average Jaccard similarity of their encounter and friend set of 0.112, with only 8, 278 users with a nonzero value. In general, the sets of encounters and of friends of a user can significantly different and often have empty intersection. Despite epidemic processes spreading on the friendship and on the encounter network evolve in a qualitatively similar way, the differences in local connectivity determined by the two definitions of edges might result in very different sets of nodes at risk of infection.

2.3 The static encounter network

To argue that our results are not driven by the static nature of the friendship network as opposed to the time-varying nature of the encounter network, we also consider a static version of the encounter network. Let $E_u \;\; \cup_{t \in T} E_u t \subseteq U$ be the set of encounters of u during T, and $E \;\; \cup_{t \in T} \cup_{u \in U} \{u, v : v \in Et\} \subseteq U \times U$ be the set of encounters between users in U. The *static encounter network* is $N_E \;\; U, E$. We restrict our attention to the giant component of the static encounter network, which includes 113, 187 users (whereas the second largest component has 23 users).

3 INFECTION DYNAMICS

To model the spread of an infectious disease, we consider a Susceptible-Infected (SI) process [2] , in which nodes never recover after being infected. Here, we give a general definition of the process that applies to both the static and the time-varying networks defined above. Given a set of nodes $\mathcal{V}$, a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ and a set of time indices $\mathcal{T}$, let $\{Nt\}_{t \in \mathcal{T}}$ be a sequence of networks, where $Nt \;\; \mathcal{V}, \mathcal{E}t$ with $\mathcal{E}t \subseteq \mathcal{E}$. For a static network, $\mathcal{E}t \;\; \mathcal{E}$ for all of $t \in \mathcal{T}$.

Let $\mathcal{I}t$ denote the set of infected nodes at time t, of cardinality It. The infection starts at time $t \;\; 0$ from a set $\mathcal{I}0$ of infected seeds.

Consider any $t > 0$. The infection spreads from the set of already infected nodes $\mathcal{I}t - 1$ as follows. For each non-infected node $v \in \mathcal{V} \backslash \mathcal{I}t - 1$, let $d_v t \;\; |\{u \in \mathcal{I}t - 1 : u, v \in \mathcal{E}t\}|$, that is, the number of neighbors of v at time t which are infected at time $t - 1$. Let $Bt \;\; \{v \in \mathcal{V} \backslash \mathcal{I}t - 1 : d_v t > 0\}$, that is, the set of susceptible nodes at time t. Each node $v \in Bt$ gets infected with probability $\beta d_v t$, where $\beta \in [0, 1]$ is the rate of infection.

When $\beta \;\; 1$ the infection process is deterministic and, at time t, all non-infected neighbors of the nodes infected by time $t - 1$ become infected. For finite values of β, the infection spreads in a stochastic way. We consider different values of β for the different networks, due to their different connectivity ($\beta \;\; 0.5$ on the encounter network, and $\beta \;\; 0.01$ on the static networks, unless differently stated).

For the time-varying networks defined above (i.e., the encounter network and the time-varying friendship network), $\mathcal{T} \;\; T$. The infection will propagate for $|T|$ time steps, resulting in an infected population $\mathcal{I}|T|$. For static networks (i.e., the friendship network and static encounter network), $\mathcal{T} \;\; [0, \infty$ and the infection propagates until $\mathcal{I}t \;\; \mathcal{V}$ (i.e., until the entire population is infected).

3.1 Infection time

Given a realization of the infection process, for each $m \in [0, |\mathcal{V}|]$, let

$$tm \;\; \min\{t : It \geq m\}.$$

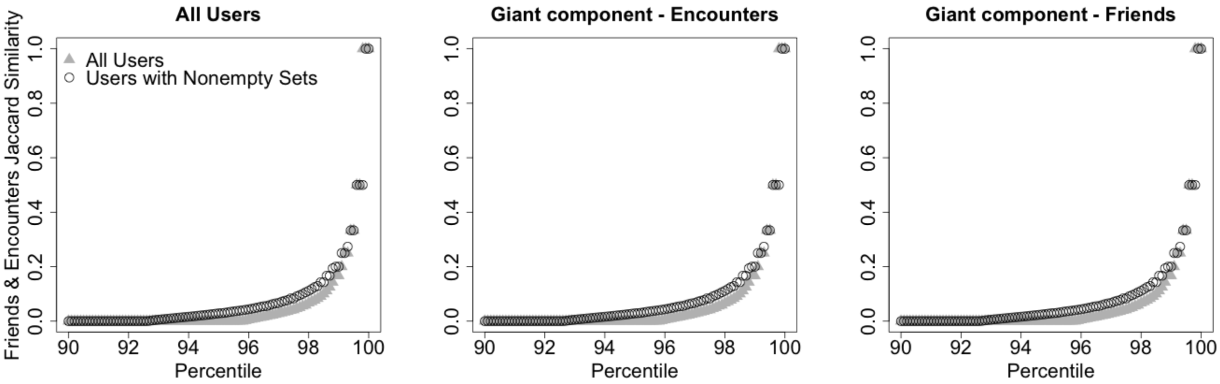

Figure 1: Percentile plot of the Jaccard similarity of the set all user's friends and the set of all user's encounters. Left: all non-singleton users. Center: users in the giant component of the network defined by all encounters. Right: users in the giant component of the network defined by all friendship ties.

The random variable tm denotes the first time in which at least m nodes are infected. Given a realization of the SI process on a time-varying network, let $tm \quad \infty$ for $M > \mathcal{I}|T|$. In what follows, the notation $t_A m$ indicates that nodes on a specific network A are considered (e.g. A can be the friendship or the encounter network, even if the infection spreads on the encounter network).

3.2 Seed selection

In a static network, seeds are chosen at random and without replacement. In a time-varying network, the infection can start propagating at the first time t in which there is an edge between an infected seed and a non-infected node, that is, at time

$$t_0\mathcal{I}0 \quad \min\{t : \exists u, v \in \mathcal{E}t \text{ for some } u \in \mathcal{I}0, v \in \mathcal{V}\backslash\mathcal{I}0\}.$$

As a remark, for $\beta < 1$, it is possible that no node is infected at time t_0. Seeds are selected uniformly at random and without replacement among all nodes $v \in \mathcal{V}$ such that $t_0\{v\} \leq 500$, that is, nodes that have a neighbor in the time-varying network by time $t \quad 500$.

3.3 Real infection and predicted infection

Assuming simulated infections on the time-varying encounter network as the benchmark, we quantify the extent to which the friendship network provides good enough prediction of risk at the individual level. Simulated infections on the static version of the encounter network will serve instead as a comparison, in order to characterize how the loss of temporal information affects prediction accuracy. In other words, we consider infection dynamics on the encounter network as the real infections, and try to predict them by running infection dynamics on the friendship network and on the static version of the encounter network.

4 EPIDEMIC RISK AND NETWORK DISTANCE

In this section we show that distance on the friendship network is correlated to epidemic rick. Given and infection initiated at a single seed and spreading on the encounter network, nodes at a shorter distance from the seed on the encounter network have a higher probability of becoming infected. In the rest of the section, we always consider infections spreading on the encounter network and distance defined on the friendship network.

Given nodes s and s' in the friendship network, let ds, s' denote their distance (i.e., the length of the shortest path connecting them). Given node s and an integer $d > 0$, let

$$N_d s \quad \{s : ds, s' \quad d\}$$

be the set of nodes at distance d from s, and let $n_d s$ be its cardinality. $N_1 s$ and $n_d s$ denote the set of neighbors and the degree of s, respectively.

Let i denote an infection process, and s_i the selected seed. Given an infection initiated at a seed s_i until time T, let $\mathcal{I}s_i$ be the set of infected nodes at time T. For each $d > 0$ let

$$\mathcal{I}_d s_i \quad \mathcal{I}s_i \cap N_d s$$

be the set of infected nodes that are at distance d from s_i on the encounter network. The infection fraction of nodes at distance d from s_i is defined as

$$r_d s_i \quad \frac{|\mathcal{I}_d s_i|}{n_d s}.$$

The empirical average of $r_d s_i$ over S simulations is given by

$$\bar{r}_d \quad \frac{1}{S}\sum_{i1}^{S} r_d s_i,$$

and represents the risk of becoming infected if the seed is at distance d.

As the spreading of an infection process depends on the infection rate β, we write $\bar{r}_d\beta$ to compare infection processes with different infection rate. Given a node s in the encounter network, we recall that $t_0\{s\}$ is the first time period in which s has an edge (that is, the smallest t such that $E_u t > 0$). As we consider infections spreading on the encounter network and distance on the friendship network, we consider seeds that are present in both networks. In each simulation, a single seed is selected uniformly at random between all nodes $s \in |U_F \cap U_E|$ such that $t_0\{s\} \leq 500$ (as infections on time-varying

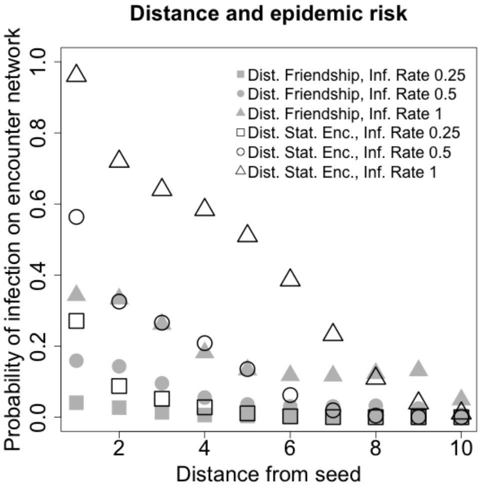

Figure 2: Risk of infection on the encounter network versus distance from the infection seed. The y-axis plots the empirical probability that nodes at distance d become infected on the encounter network (distance on the friendship network: grey; distance on the static encounter network: white).

networks spread for a limited number of time steps, we require them to start early enough). For each $\beta \in \{0, 0.1, 0.25, 0.5\}$ we run 10,000 simulations. The empirical estimates of $\bar{r}_d\beta$ for $1 \leq d \leq 10$ are shown in Figure 2 and Table 1.

5 PREDICTIVE ACCURACY OF THE FRIENDSHIP NETWORK

In order to evaluate how accurately the friendship network predicts epidemic risk at a microscopic level, we consider infection processes initiated at the same seed and spreading independently of each other, and compare the sets of infected nodes.The unpredictability of epidemic risk is due to the structural differences of the different networks as well as to the randomness of the infection processes. Therefore, for each of 5,000 (a node can be selected multiple times as the seed), we consider four infection processes: two infection processes on the encounter network that spread independently of each other, 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 $A \in \{E_1, E_2, F, S\}$ initiated at seed s, let $t_Am; s$ be the first time at which at least m nodes are infected (the quantity might be undefined if the infection does not reach at least m individuals). When $t_Am; s$ is defined, let $I_Am; s$ be the corresponding infected set (of size at least m). Consider two infections $A, B \in \{E_1, E_2, F, S\}$, $A \neq B$, initiated at the same seed s and either spreading on two different networks, or spreading on the same network but independently of each other. Given a target m, if both $I_Am; s$ and $I_Bm; s$ are defined, their Jaccard similarity is given by

$$J_{A,B}m; s \quad \frac{|I_Am; s \cap I_Bm; s|}{|I_Am; s \cup I_Bm; s|},$$

where, given a set X, $|X|$ denotes its cardinality. These measures allow to characterize how accurately the friendship network and the static version of the encounter network predict epidemic risk on the encounter network (in the SI, we also consider the precision metrics $|I_Am; s \cap I_Bm; s|/|I_Am; s|$ and $|I_Am; s \cap I_Bm; s|/|I_Bm; s|$, which provide similar observations and results).

Results are shown in Figure 3 and Table 2. Starting from the left, the first panel plots the metrics $J_{E_1,E_2}m; s$ for all seed selections (and a range of values of the target infection size m), and represents the baseline unpredictability due solely to the randomness of processes initiated at the same seed and spreading independently on the encounter network. The second panel shows the metrics $J_{E_1,S}m; s$, which includes the unpredictability due to the loss of temporal information in the static version of the encounter network. The third panel shows the metrics $J_{E_1,F}m; s$, which represents the unpredictability of using the friendship network to predict risk on the encounter network. The fourth and rightmost panel shows the Jaccard similarity for pairs of random sets of m nodes connected on encounter network. Such metric represents what is achievable by random guessing, without the knowledge of either the friendship or the encounter network. Higher values of the y-axis correspond to higher prediction accuracy. For each value of the target m separately, $J_{E_1,E_2}m; s$ has larger average than both $J_{E_1,S}m; s$ and $J_{E_1,F}m; s$, and that $J_{E_1,S}m; s$ has larger average than $J_{E_1,F}m; s$. Notably, the intersections of the infected sets on the friendship and encounter networks are substantially and significantly larger than the intersection of random sets (average Jaccard similarity $1.2 \cdot 10^{-2}$ vs. $8.3 \cdot 10^{-4}$, two-sample t-tests, p-value$< 2.2 \cdot 10^{-16}$). This shows the value of using the friendship network for predicting epidemics risk even if the infection is driven by physical encounter.

Together, the similarity measures $J_{\cdot,\cdot}m; s$ allow to characterize how the randomness of the infection process, the temporal ordering of the encounters and the structural differences between the networks affect the predictability of epidemic risk. Our analyses show that friendship helps identifying the individuals at risk of infection even if the epidemic is driven by physical encounter (compare the third and fourth panels of Figure 3). This is an important result, as in practice it might be feasible to track friendship or other forms of static relationship, but infeasible to track or predict physical encounter. However, knowledge of the friendship network does not allow to reach the same accuracy as knowing the encounter network (which is usually unavailable or extremely costly to get). On the one hand, the randomness of the infection determines unpredictability of the set of infected individuals, even between independent processes spreading on the encounter network and initiated at the same seed (first panel of Figure 3). On the other hand, structural differences amplify such unpredictability when comparing processes spreading on the friendship network and the encounter network (first and third panels of Figure 3). In addition, our results are not only driven by the static nature of the friendship network opposed to the time-varying nature of the encounter network, as the static version of the encounter network provides more accurate prediction of risk than the friendship network (second and third panels of Figure 3).

Table 1: Epidemic risk with respect to distance on the friendship network.

β	$\bar{r}_1\beta$	$\bar{r}_2\beta$	$\bar{r}_3\beta$	$\bar{r}_4\beta$	$\bar{r}_5\beta$	$\bar{r}_6\beta$	$\bar{r}_7\beta$	$\bar{r}_8\beta$	$\bar{r}_9\beta$	$\bar{r}_{10}\beta$
0.10	$3.9{\cdot}10^{-3}$	$7.1{\cdot}10^{-4}$	$2.1{\cdot}10^{-4}$	$7.01{\cdot}10^{-5}$	$3.2{\cdot}10^{-5}$	$2.3{\cdot}10^{-5}$	$1.3{\cdot}10^{-5}$	$2.1{\cdot}10^{-5}$	$1.1{\cdot}10^{-5}$	0
0.25	0.041	0.027	0.014	0.006	0.003	0.003	0.002	0.003	0.001	$1.6{\cdot}10^{-4}$
0.50	0.159	0.143	0.095	0.055	0.036	0.031	0.030	0.032	0.025	0.007
1.00	0.343	0.333	0.262	0.182	0.133	0.118	0.116	0.123	0.131	0.049

Table 2: Single seed infection on the encounter network and the static (encounter and friendship) networks. Stochastic infection - Similarity measures.

m	$\langle J_{E_1^t,E_2^t} m\rangle$	$\langle J_{E_1^t,E_1^s} m\rangle$	$\langle J_{E_1^t,F_1} m\rangle$	$\langle \bar{J}_{E_1^t,E_2^t} m\rangle$	$\langle \bar{J}_{E_1^t,E_1^s} m\rangle$	$\langle \bar{J}_{E_1^t,F_1} m\rangle$
500	0.115	0.039	0.012	0.270	0.315	0.323
1000	0.159	0.056	0.019	0.325	0.561	0.454
2000	0.220	0.082	0.031	0.438	0.716	0.615
5000	0.316	0.129	0.056	0.571	0.776	0.744
10000	0.397	0.178	0.081	0.664	0.790	0.806
20000	0.466	0.249	0.110	0.788	0.835	0.830

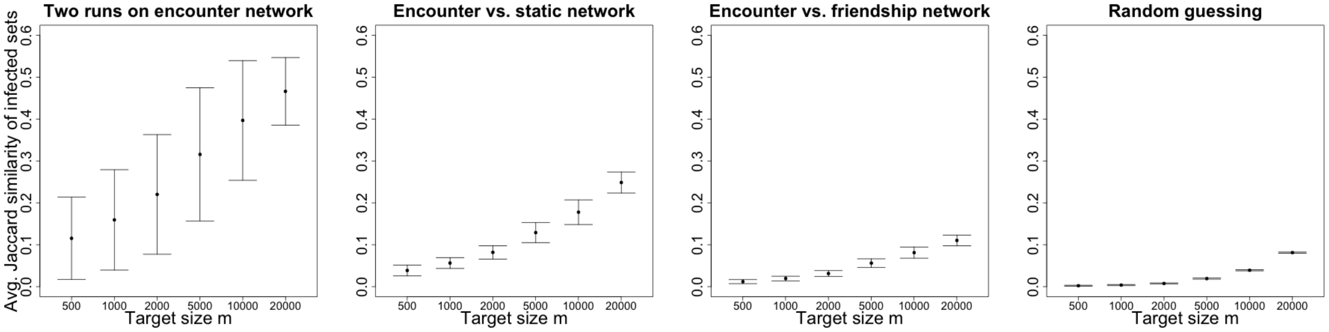

Figure 3: Predictability of nodes' epidemic risk. First panel: $J_{E_1,E_2}m; s$. Second panel: $J_{E_1,S}m; s$. Third panel: $J_{E_1,F}m; s$. Fourth panel: pairs of random infected sets of size m.Black points represents averages, bars standard deviations. Higher values of the y-axis correspond to higher prediction accuracy.

6 PERIODICAL MONITORING AND PREDICTION

In addition to the predictive power that knowledge of the friendship network brings on its own, here we show how periodical, even if relatively infrequent, monitoring of the infected population can boost the prediction capabilities of the friendship network. In particular, we show that periodical monitoring of the benchmark infection spreading on the encounter network allows to correct the predicted infection spreading on the friendship network, substantially increasing accuracy. This corresponds to a scenario in which the investigator has knowledge of the friendship network but, in addition, is able to observe the infected population at fixed intervals. Periodical reports of the infection are usually available in the case of real epidemics (e.g., weekly or monthly). After each observation, the set of infected individuals according to the dynamics on the friendship network is updated to match the set of infected individuals according to the dynamics on the encounter network.

Given a seed s connected on both the encounter and the friendship networks (and such that $t_0 s \leq 900$), we consider an infection spreading on the encounter network and one spreading on the friendship network with periodic corrections (denoted by F and E respectively), for 500 time steps each and independently of each other. Given an observation window W, every W time steps the predicted infected set $I_F kW$ on the friendship network is corrected to match the benchmark infected set $I_E kW$ on the encounter network. That is,

$$I_F kW \quad I_E kW, \text{ for each } k > 0.$$

and between time kW and $k \;\; 1W - 1$ the set predicted infected $I_F t$ grows according to the ties of the friendship network (because the encounter network, driving the real infection, is not known). We are interested in comparing the sets $\mathcal{I}_E t$ and $I_F t$ at times $t \;\; kW - 1$, that is, right before each correction. Let

$$J_{E,F} k; s, W \quad \frac{I_E kW - 1 \cap I_F kW - 1}{I_E kW - 1 \cup I_F kW - 1},$$

be the Jaccard similarity of the infected sets on the two networks right before a correction (the notation shows its dependence on W

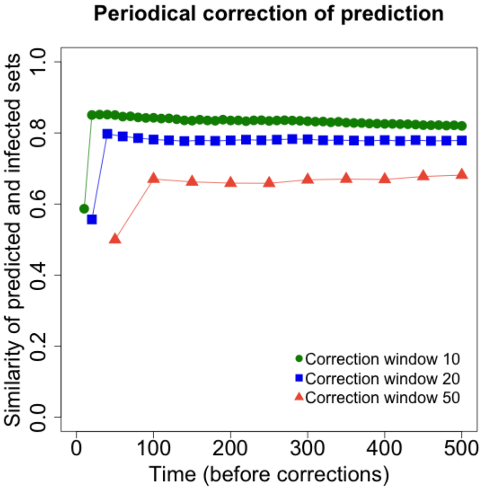

Figure 4: Average Jaccard similarity between the predicted infected sets (according to infections spreading on the friendship network, and periodic corrections) and the infected sets given by infections spreading on the encounter network. Higher values of the y-axis correspond to higher prediction accuracy.

and on the realization of the infection process, represented by its seed s).

Figure 4 plots the average Jaccard similarity of the sets of all infected individuals in the two processes right before each correction (times $kW-1$, including all previous updates of the predicted infected sets), for window length $W \in \{10, 20, 50\}$ (6000 simulations for each W). Note that, as each infection process is run for T 500 time steps, the number of corrections (and therefore the number of points in the plots) depends on the choice of W and equals T/W. A high level of prediction accuracy is established early in the process (after the first correction) and maintained over time. The accuracy decreases with larger window size, but even W 50 guarantees good accuracy. Our results suggest that the ability to periodically monitor who is infected (according to the infection on the encounter network) is key to overcome the limits of the friendship network in predicting epidemic risk.

In order to compare all window sizes $W \in \{10, 20, 50\}$, we consider all time steps corresponding to a correction for all choices of W and ignore the first correction (i.e., we consider times $100k$ for $1 \leq k \leq 5$). The trend of the average of the Jaccard similarity $J_{E,F}k; s, W$ with respect to time t and window size W is captured by a linear relationship. The measure is lower in the case of W 50 (-0.188 with respect to W 10, p-value $2.74 \cdot 10^{-10}$), value for which it increases over time ($3.29 \cdot 10^{-3}$ every 100 time steps, p-value $1.27 \cdot 10^{-3}$).

7 TARGETED IMMUNIZATION

In addition to analyzing the power of the online friendship network for real time monitoring during the response phase of an epidemic, we consider as well how it can improve preparedness through immunization campaigns, which can take the form of physical vaccination or information campaigns informing and advocating for safe practices. In this sense, the friendship paradox (i.e., the average friend of an individual is more connected than the average individual [10]) has shown that name-a-friend methods improve the prediction of the peak of an epidemic outbreak [6] and the spread of information online [12]. Instead of considering a scenario in which the same network defines both social ties and infection, here we show that such policies can be effective when social ties are defined according to an online friendship network and infection spreads on an encounter network.

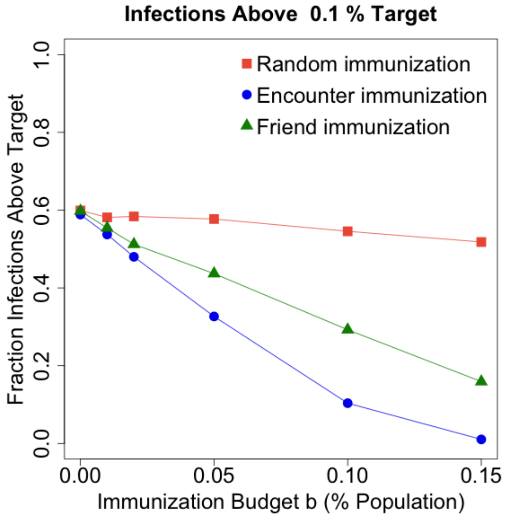

Figure 5: Fraction of infections that do not die out in their early stages as a function of the immunization budget b and the immunization method. Lower values of the y-axis correspond to more effective immunization strategies.

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 spreading 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). The method results in a more effective use of the immunization budget, substantially increasing the probability that an infection dies out in its early stages (Figure 5) and strongly reducing the final infection size (Figures 6) with respect to random immunization. Moreover, it only requires a small additional cost (in terms of the number of immunized individuals) to obtain the same effect as an ideal strategy that targets future encounters rather than friends (encounter immunization).

Immunization budget is expressed as a fraction b of the entire population. For $b \in \{1\%, 2\%, 5\%, 10\%, 15\%\}$, Figure 5 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 value of b). We consider a 0.1% target for the final infection as an indicator that the infection did not die out in its early stages. Lower values of the y-axis correspond to more effective immunization strategies. The trend in Figure 5 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 of 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 $4.03 \cdot 10^8$). Despite its simplicity, friend immunization provides comparable effectiveness as the encounter immunization strategy (which would require knowledge of the encounter network), at a small additional cost. For a fixed value of the y-axis, observe the limited extra immunization budget required to reach that performance employing friend immunization rather than encounter immunization.

Regarding the size of the infected population, Figure 6 shows (for b 5%) the fraction of infections that reach given target sizes (5%, 10%, 15% of the entire population in the left, middle and right panel respectively) as a function of the infection start time t_0s of the seed for all immunization methods (5000 simulations for each immunization method). The y-axis shows the fraction of infections whose final size is above the given target, and lower values correspond to more effective immunization strategies. Friend immunization provides a large advantage with respect to random immunization, and unlike the latter, its effectiveness increases with increasing immunization budget.

8 CONCLUSION

Established research suggests that biological networks, social networks and the Internet are governed by similar rules [1, 3, 34]. Similar models have been proposed to characterize the spread of epidemics, information, behaviors, and cultural norms. In line with this results, our work shows that the friendship network provides useful information for identifying the individuals at risk even if the infection spreads on the encounter network. However, due to the structural differences between the two networks, accuracy does not come close to the ideal case in which the encounter network is known. Even though the two networks are different, our simulations show that knowledge of the friendship network enables effective monitoring and immunization strategies. Very high prediction accuracy using the friendship network can be reached and maintained if periodical yet infrequent reports of the infected population are available. 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, and requires a small additional investment to reach the same performance as knowing the encounter network.

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. Taken together, our results highlight the opportunity of using a friendship network for predicting, monitoring and containing epidemics. In real scenarios, friendship, family or professional networks (which can be considered static or almost static) are more likely to be available than time-varying networks of physical encounter, which would require extensive tracking of the population. In addition, the encounter network is fully accessible only in a context of "prediction in retrospect", as in the case of the present work. 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 [5, 14, 40, 42]. From a practical perspective, networks based on social relationships (such as a friendship network) might be complemented by information about past encounter.

We considered reviews as a proxy of physical encounter – an edge is active between two users on day t if they both posted a review to or a tip about the same business on day t. This constitutes an approximation of real physical encounter, which would require users to visit (rather than review) a business at about the same time. In order to justify this assumption, we observe that the time of a review or tip is a proxy of the time of the visit to a business, and that infections do not necessarily require direct physical contact. In fact, in the case of certain airborne diseases, particles can remain suspended in the air for several hours after an infected individual has been in a room [4]. In the context of our dataset, after an infected user visits a business, the infection might spread to customers who visit the business later in the day. Other proxies of physical encounter, such as proximity measured by Bluetooth devices, are usually limited to small population, and suffer different limitations (e.g., the signal passes through walls).

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 includes more than 100k individuals and spans more than 4y 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

[1] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics* 74, 1 (2002), 47.

[2] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious diseases of humans: dynamics and control*. Vol. 1. Wiley Online Library.

[3] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. 2006. Complex networks: Structure and dynamics. *Physics reports* 424, 4 (2006), 175–308.

[4] Gabrielle Brankston, Leah Gitterman, Zahir Hirji, Camille Lemieux, and Michael Gardam. 2007. Transmission of influenza A in human beings. *The Lancet infectious diseases* 7, 4 (2007), 257–265.

[5] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. 2006. The scaling laws of human travel. *arXiv preprint cond-mat/0605511* (2006).

[6] Nicholas A Christakis and James H Fowler. 2010. Social network sensors for early detection of contagious outbreaks. *PloS one* 5, 9 (2010), e12948.

[7] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.

[8] Odo Diekmann and Johan Andre Peter Heesterbeek. 2000. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*. Vol. 5. John Wiley & Sons.

[9] Katayoun Farrahi, Remi Emonet, and Manuel Cebrian. 2014. Epidemic contact tracing via communication traces. *PloS one* 9, 5 (2014), e95133.

[10] Scott L Feld. 1991. Why your friends have more friends than you do. *Amer. J. Sociology* 96, 6 (1991), 1464–1477.

[11] Neil M Ferguson, Derek AT Cummings, Simon Cauchemez, Christophe Fraser, et al. 2005. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437, 7056 (2005), 209.

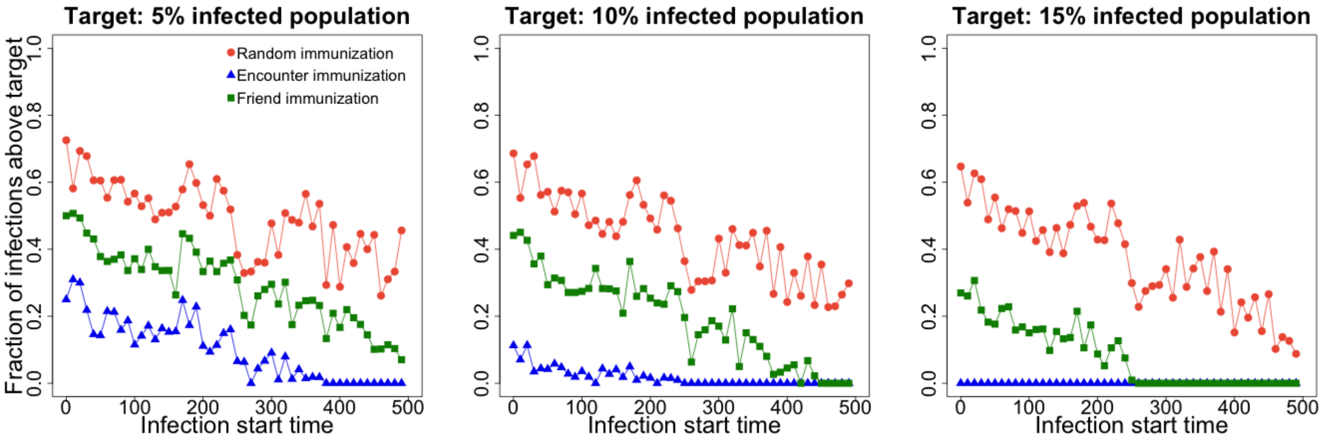

Figure 6: Final infection size as a function of the immunization method and the infection start time, given immunization budget b 5% of the entire population. Lower values of the y-axis correspond to more effective immunization strategies.

[12] Manuel Garcia-Herranz, Esteban Moro, Manuel Cebrian, Nicholas A Christakis, and James H Fowler. 2014. Using friends as sensors to detect global-scale contagious outbreaks. *PloS one* 9, 4 (2014), e92413.

[13] Marcelo FC Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis Chao, Ira Longini, M Elizabeth Halloran, and Alessandro Vespignani. 2014. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. *PLoS currents* 6 (2014).

[14] Marta C Gonzalez, Cesar A Hidalgo, and A-L Barabasi. 2008. Understanding individual human mobility patterns. *arXiv preprint arXiv:0806.1256* (2008).

[15] Thilo Gross and Bernd Blasius. 2008. Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface* 5, 20 (2008), 259–271.

[16] M Elizabeth Halloran, Ira M Longini, Azhar Nizam, and Yang Yang. 2002. Containing bioterrorist smallpox. *Science* 298, 5597 (2002), 1428–1432.

[17] Till Hoffmann, Mason A Porter, and Renaud Lambiotte. 2012. Generalized master equations for non-Poisson dynamics on networks. *Physical Review E* 86, 4 (2012), 046102.

[18] Petter Holme. 2005. Network reachability of real-world contact sequences. *Physical Review E* 71, 4 (2005), 046119.

[19] Petter Holme. 2015. Information content of contact-pattern representations and predictability of epidemic outbreaks. *Scientific reports* 5 (2015).

[20] Lars Hufnagel, Dirk Brockmann, and Theo Geisel. 2004. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America* 101, 42 (2004), 15124–15129.

[21] Hang-Hyun Jo, Juan I Perotti, Kimmo Kaski, and János Kertész. 2014. Analytically solvable model of spreading dynamics with non-Poissonian processes. *Physical Review X* 4, 1 (2014), 011041.

[22] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. 2011. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E* 83, 2 (2011), 025102.

[23] Matt J Keeling and Pejman Rohani. 2008. *Modeling infectious diseases in humans and animals*. Princeton University Press.

[24] Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, Jari Saramäki, and Márton Karsai. 2012. Multiscale analysis of spreading in a large communication network. *Journal of Statistical Mechanics: Theory and Experiment* 2012, 03 (2012), P03005.

[25] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. *Pervasive Computing* (2009), 176–183.

[26] Jim Koopman. 2004. Modeling infection transmission. *Annu. Rev. Public Health* 25 (2004), 303–326.

[27] Kai Kupferschmidt. 2014. Estimating the Ebola epidemic. *Science* 345, 6201 (2014), 1108–1108.

[28] Scott Lederer, Jennifer Mankoff, and Anind K Dey. 2003. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03 extended abstracts on Human factors in computing systems*. ACM, 724–725.

[29] Suyu Liu, Nicola Perra, Márton Karsai, and Alessandro Vespignani. 2014. Controlling contagion processes in activity driven networks. *Physical review letters* 112, 11 (2014), 118702.

[30] Alun L Lloyd and Robert M May. 2001. How viruses spread among computers and people. *Science* 292, 5520 (2001), 1316–1317.

[31] Eric T Lofgren, M Elizabeth Halloran, Caitlin M Rivers, John M Drake, Travis C Porco, Bryan Lewis, Wan Yang, Alessandro Vespignani, Jeffrey Shaman, Joseph NS Eisenberg, et al. 2014. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences* 111, 51 (2014), 18095–18096.

[32] Ira M Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek AT Cummings, and M Elizabeth Halloran. 2005. Containing pandemic influenza at the source. *Science* 309, 5737 (2005), 1083–1087.

[33] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. 2015. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one* 10, 9 (2015), e0136497.

[34] Mark EJ Newman. 2003. The structure and function of complex networks. *SIAM review* 45, 2 (2003), 167–256.

[35] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical review letters* 86, 14 (2001), 3200.

[36] Nicola Perra, Andrea Baronchelli, Delia Mocanu, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2012. Random walks and search in time-varying networks. *Physical review letters* 109, 23 (2012), 238701.

[37] Anatol Rapoport. 1953. Spread of information through a population with sociostructural bias: I. Assumption of transitivity. *Bulletin of Mathematical Biology* 15, 4 (1953), 523–533.

[38] Luis EC Rocha, Fredrik Liljeros, and Petter Holme. 2011. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology* 7, 3 (2011), e1001109.

[39] Hiroki Sayama, Irene Pestov, Jeffrey Schmidt, Benjamin James Bush, Chun Wong, Junichi Yamanoi, and Thilo Gross. 2013. Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications* 65, 10 (2013), 1645–1664.

[40] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[41] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, et al. 2011. Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC medicine* 9, 1 (2011), 87.

[42] Lijun Sun, Kay W Axhausen, Der-Horng Lee, and Xianfeng Huang. 2013. Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences* 110, 34 (2013), 13774–13779.

[43] Thomas W Valente. 1995. Network models of the diffusion of innovations. (1995).

[44] Alexei Vazquez, Balazs Racz, Andras Lukacs, and Albert-Laszlo Barabasi. 2007. Impact of non-Poissonian activity patterns on spreading processes. *Physical review letters* 98, 15 (2007), 158702.