Contagion and ranking processes in complex networks: the role of geography and interaction strength

PhD Thesis Proposal

Qian Zhang
College of Computer and Information Science
Northeastern University, Boston, MA

January 14, 2014

Abstract

The recent global surge in the use of technologies such as social media, smart phones and GPS-enabled devices has provided abundant resources to understand dynamical processes on complex networks and a unique chance to characterize geospatial and temporal distribution of real time social events. Meanwhile, the easy accessibility to bibliographic data and geographical database allow better understanding of scholarly networks and in charting the creation of knowledge globally. Moreover, the availability of large-scale communication datasets presents new opportunities to study information dissemination on social networks. In this thesis we focus on the diffusion processes on complex networks aggregated from these data. First, we investigate geospatial and temporal features of a publication dataset. We characterize the knowledge diffusion pattern between worldwide urban areas and its temporal evolution, and identify the key cities in the scientific research in physics. Second, we propose to detect and predict seasonal flu epidemics in countries of interest with geolocalized Twitter signals. In the early stage of a flu season, tweets containing information related with influenza-like illness indicate the spatial distribution of possible initial infected cases. Modeling disease and spreading dynamics with these initial seeds can provide the possibility of forecasting ILI cases in the coming month. Beyond geospatial information, we also investigate the role of other network properties play on information diffusion process. For a human-to-human communication network, we survey different definitions of weak ties and develop a novel link property *importance* to characterize the strength of ties. Controlling weak ties defined by *importance* can more efficiently confine the information spreading within a small community than controlling weak ties under other definitions. Last but not least, we find a phase transition between absorbing and active states of the classic Maki-Thompson rumor spreading model on random networks. The parameters of the contagion process as well as the network architecture determine whether the rumor will spread globally or whether it will be confined within a small neighborhood.

1 Introduction

Digital records of social interactions and publication datasets have been increasingly accessible in the past decade. These high resolution and large-scale datasets obtained through Internet, mobile devices and pervasive technologies have enabled a wealth of research in dynamical processes occurring on top of complex networks [71, 11]. Digitalization of bibliographic data help understanding the geospatial distribution of millions of publications, and citations at different granularities [12, 51, 42, 33, 64, 56]. Considering the geographical spread of human infectious diseases, data-driven computational models [9] have been proven to be indispensable tools in guiding public health policies by the availability of human mobility datasets at large scale [25, 7]. Moreover, these data have enabled the theoretical understanding of critical phenomena of nonequilibrium phase transitions in infectious disease spread mediated by complex network structures and mobility schemes [26, 10]. Similarly, accessibility of highly detailed mobile phone communication datasets at large scale [62, 50, 48, 18, 49] provides unique opportunities for studying information diffusion in real social systems from both theoretical and computational points of view. Furthermore, online social network platforms such as Twitter provide abundant sources of real time events. Each user on such the platform can be considered as a social sensor and each tweet as sensory information [67]. These social sensors provide a unique chance to detect and predict real time social events such as earthquake [67] flu [27, 3], election [70], reality-singing competition [24]. In this thesis work, we take advantage of such large amount of data and propose detailed studies on the knowledge diffusion on the citation networks geolocalized at the level of urban areas; detecting and predicting seasonal flu with geolocalized Tweets as initial seeds; measuring the strength of weak ties and their role on diffusion process on a large scale human-to-human communication networks; and numerical

and analytical understanding the phase transition of a classic rumor spreading model on random networks.

Geospatial distribution of knowledge production and consumption in Physics. We analyze the entire publication database of the American Physical Society generating longitudinal (50 years) citation networks geolocalized at the level of single urban areas. We define the knowledge diffusion proxy, and scientific production ranking algorithms to capture the spatio-temporal dynamics of Physics knowledge worldwide. By using the knowledge diffusion proxy we identify the key cities in the production and consumption of knowledge in Physics as a function of time. The results from the scientific production ranking algorithm allow us to characterize the top cities for scholarly research in Physics.

Detecting and predicting seasonal flu using Twitter data. We consider geolocalized tweets containing ILI related keywords as social sensory information indicating seasonal flu cases in the early stage of a flu season. We calibrate such sensory information with surveillance data as well as census population data and generate initial infectious seeds for each urban area in a given country of interests. The initial seeds fuse into a stochastic epidemic model, GLEAM (global epidemic and mobility model) [9, 7], which considers detailed disease dynamics inside a single urban area and transmission dynamics between different urban areas. A large amount of stochastic simulations in a sampled parameter space provide a pool of candidate data, which describe possible number of ILI cases as a function of time. We use the known surveillance data in the current and past seasons to find the best candidate that represents the real scenario of the seasonal flu, and generate predictions of the number of cases in the coming time window.

Strength of weak ties on diffusion process on mobile communication networks. Based on Granovetter's definition, weak ties in a social network are responsible for the information transmission through otherwise disconnected communities because they are on the shortest path between many nodes [38]. Sticking to Granovetter's definition, we consider both the role each link plays on information diffusion process and its topological role on the network. We take advantage of collections of human-to-human communication records in real life, and define a novel link property called *importance* to quantitatively characterize the significance of a link in the diffusion process. We investigate the structural roles of weak ties and the effect of weakening weak ties on the diffusion control under different definitions of strength.

Phase transition of rumor spreading process on complex networks. We report simulation and analytical results showing that, unlike the past studies showed, there exists a phase transition between absorbing and active states for Maki-Thompson rumor model on uncorrelated random networks. The parameters of the contagion process as well as the network architecture determine whether the rumor spreads globally or is absorbed by a small neighborhood of the initial spreader.

2 Geospatial distribution of knowledge production and consumption in Physics

The digitalization of publication data has propelled bibliographic studies allowing for the first time access to the geospatial distribution of millions of publications, and citations at different granularities [58, 32, 55, 12, 51, 42, 33, 64, 56]. More precisely, authors' name, affiliations, addresses, and references can be aggregated at different scales, and used to characterize publications and citations patterns of single papers [65, 22], journals [34, 13], authors [40, 31, 41], institutions [15], cities [16], or countries [47]. Such large databases are extremely useful in charting the creation of knowledge, they are also pointing out the limits of our conceptual and in deep understanding of the dynamics ruling the diffusion and fruition of knowledge across the the social and geographical space.

In this project we focus on database of articles published in the American Physical Society (APS) journals in a fifty-year time interval (1960-2009) [6]. For each paper we geolocalize the institutions contained in the authors' affiliations and associate each paper in the database with specific urban areas. In order to geolocalize the articles, we first process each affiliation string and try to match country or US state names from a list of known names and their variations in different languages. We crosscheck the results with Google Map API obtaining validated location information for 97.7% of affiliation strings,

corresponding to $445,223$ articles. It is worth noticing that we do not use Google Map API (or other map APIs like Yahoo! or Bing) directly for geocoding because, to our best knowledge, there are no accuracy guarantees to these API results. For each affiliation string with an extracted country or US state name, we also match the city name against GeoName database [35] corresponding to its country or US state. 92.6% of affiliation strings with extracted city names are subsequently verified with Google Map API. Finally, a total of $425,233$ publication articles successfully pass the filters we describe here.

In order to construct the geolocalized citation network we consider nodes (urban areas) and directed links representing the presence of citations from a paper with affiliation in one urban area to a paper with affiliation in another urban area. For example, if a paper written in node i cites one paper written in node j there is an link from i to j, i.e., j receives a citation from i and i sends a citation to j. Each paper may have multiple affiliations and therefore citations have to be proportionally distributed between all the nodes of the papers. For this reason we weight each link in order to take into account the presence of multiple affiliations and multiple citations. This defines a time resolved, geolocalized citation network including 2,307 cities around the world engaged in the production of scholarly work in the area of Physics.

2.1 Characterizing knowledge production and consumption

Following previous works [15, 56] we assume that the number of given or received citations is a proxy of knowledge consumption or production, respectively. More precisely, we assume that citations are the currency traded between parties in the knowledge exchange. Nodes that receive citations export their knowledge to others. Nodes that cite other works, import knowledge from others. According to this assumption we classify nodes considering the unbalance in their trade. Specifically we define *producers* as cities that export more than they import, and *consumers* as cities that import more than they export. More precisely, we can measure the total knowledge imported by each urban area as $\sum_j w_{ij}$ and the total export as $\sum_j w_{ji}$ in a given year. The relative trade unbalance of each urban area i is $\Delta S_i = (\sum_j w_{ji} - \sum_j w_{ij})/S$, where $S = \sum_{ij} w_{ij}$ is the total number of citations worldwide. A negative or positive value of this quantity indicates if the urban area i is consumer or producer, respectively.

With definitions of knowledge producer and consumer, we propose the knowledge diffusion proxy algorithm, in order to explicitly consider the complex flow of knowledge diffusion process between producers and consumers and to capture all possible correlations and bounds between nodes that are not directly connected. This algorithm is inspired by the *dollar experiment*, originally developed to characterized the flow of money in economic networks [5]. Formally, it is a biased random walk with sources and sinks where a citation diffuses in the network. The diffusion takes place on top of the network of net trade flows. Let us define w_{ij} as the number of citation that node i gives to j and w_{ji} as the opposite flow. We can define the antisymmetric matrix $T_{ij} = w_{ij} - w_{ji}$. The network of the net trade is defined by the matrix $\mathbf{F}$ with $F_{ij} = |T_{ij}| = |T_{ji}|$ for all connected pairs (i, j) with $T_{ij} < 0$ and $F_{ij} = 0$ for all connected pairs (i, j) with $T_{ij} \geq 0$. There are two types of nodes. Producers are nodes with a positive trade unbalance $\Delta s_i = s_i^{in} - s_i^{out} = \sum_j F_{ji} - \sum_j F_{ij}$. Their strength-in is larger than their strength-out. On the other hand, consumers are nodes with a negative unbalance Δs. On top of this network a citation is injected in a producer city. The citation follows the outgoing edges with a probability proportional to their intensities, and the probability that the citation is absorbed in a consumer city j equals to $P_{\rm abs}(j) = \Delta s_j / s_j^{in}$. By repeating many times this process from each starting point (producers) we can build a matrix with elements e_{ij} that measure how many times a citation injected in the producer city i is absorbed in a city consumer j.

In Figure. 2.1-A and Figure. 2.1-B we visualize the results considering the Top four producer cities in 2009 in the USA and in Europe respectively. We show their Top ten consumers over 20 years as function of time. The size of each circle is proportional to how many times each injected citation is absorbed by that consumer. In the plot, vertical grey strips indicate that the city was not a producer during those years (e.g. Orsay in 2008). The results show that, on average, Beijing is the top consumer for all of these producers in the past 20 years. Since China registered a big economical growth and increment of research population in the early 2000, it is reasonable to assume that, thanks to this positive stimulus, many more papers were written in its capital, a dominant city for scientific research in China. However, the fast publication growth increased

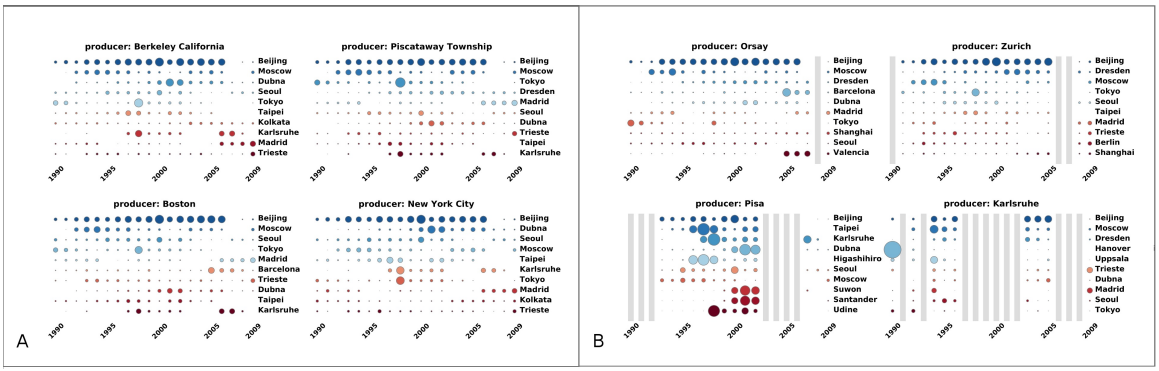


Figure 2.1: **Knowledge diffusion proxy results.** (A) The Top 4 producer cities in the USA in 2009 and their Top 10 consumers from knowledge diffusion proxy algorithm in $1990-2009$. (B) The Top 4 producer cities in the European Union 27 countries as well as Switzerland and Norway in 2009 and their Top 10 consumers from knowledge diffusion proxy algorithm in $1990-2009$. When a producer city becomes a consumer in some year, a grey strip is marked in that year. For each producer city in (A) and (B), the major consumers of the first producer city m in 20 years are plotted as a function of time from 1990 to 2009. The size of the bubble in position (Y, c) is also proportional to the counter $g_{m,c}(Y)$ in that year. The consumer cities for each producer are ordered according to the total number of counters in 20 years, i.e., $\sum_{Y_{\min}}^{Y_{\max}} g_{m,c}(Y)$.

the unbalance between sent and received citations. Each paper published in a given city imports knowledge from the cited cities. Reaching a balance might require some time. Each city needs to accumulate citations back to export its knowledge to others cities. We can speculate that in the near future cities in China might be moving among the strongest producers if a fair number of papers start receiving enough citations, which obviously depends on the quality of the research carried out in the last years. This is the case of cities like Tokyo which has gradually approached the citation balance in recent years.

2.2 Ranking scientific production

Although the knowledge diffusion proxy provides a measure of knowledge production and consumption, it may be inadequate in providing a rank of the most authoritative cities for Physics research. Indeed, a key issue in appropriately ranking the knowledge production, is that not all citations have the same weight. Citations coming from authoritative nodes are *heavier* than others coming from less important nodes, thus defining a recursive diffusion of ranking of nodes in the citation network. In order to include this element in the ranking of cities we propose the scientific production ranking algorithm. This tool, inspired by the PageRank [17], allows us to define the rank of each node, as function of time, going beyond the knowledge diffusion proxy or simple local measures as citation counts or h-index [40]. Specifically, the scientific production rank is defined for each node i according to this self-consistent equation:

$$P_i = q z_i + (1-q) \sum_j \frac{P_j}{s_j^{out}} w_{ji} + (1-q)\, z_i \sum_j P_j\, \delta\left(s_j^{out}\right). \tag{1}$$

P_i is the score of the node i, $0 \leq q \leq 1$ is the damping factor (defining the probability of random jumps reaching any other node in the network), w_{ji} is the weight of the directed connection from j to i, s_j^{out} is the strength-out of the node j and finally $\delta(x)$, is the Dirac delta function that is 0 for $x = 0$ and 1 for $x = 1$. Here we use the damping factor $q = 0.15$. The first term on the r.h.s. of Eq. (1) defines the redistribution of credits to all nodes in the network due to the random jumps in the diffusion. The second term defines the diffusion of credit through the network. Each node i will get a fraction of credit from each citing node j proportional to the ratio of the weight of link $j \to i$ and the strength-out of node j. Finally the last term defines the redistribution of credits to all the nodes in the networks due to the nodes with zero strength-out. In the original PageRank the vector $\mathbf{z}$ has all the components equal to $1/N$ (where N is the total number of nodes). Each component has the same value because the jumps are homogeneous. In this case instead, the vector $\mathbf{z}$ considers the normalized scientific

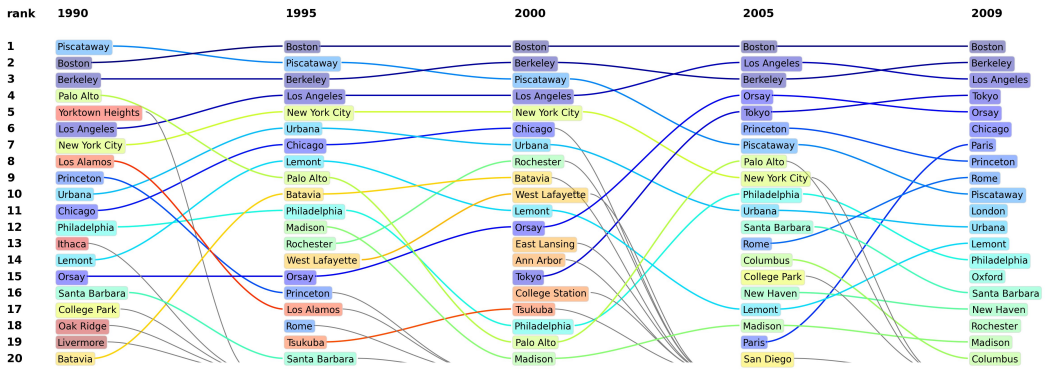


Figure 2.2: **Top 20 ranked cities as a function of time.** The plot summarizes Top 20 ranked cities in 1990, 1995, 2000, 2005 and 2009 (from left to right), and relations between the rankings in different years. The grey lines are used when the rank of that city drops out of Top 20.

credit given to the node i based on his productivity. Mathematically we have:

$$z_i = \frac{\sum_p \delta_{p,i}\ 1/n_p}{\sum_j \sum_p \delta_{p,j}\ 1/n_p} \quad , \tag{2}$$

where p defines the generic paper and n_p the number of nodes who have written the paper. It is important to notice that $\delta_{p,i} = 1$ only if the i-th node wrote the paper p, otherwise it equals zero.

As stated above, in this algorithm, the credits diffuse following citations links self-consistently, implying that not all links have the same importance. Any city in the network will be more prominent in rank if it receives citations from high-rank sources. This process ensures that the rank of each city is self-consistently determined not just by the raw number of citations but also if the citations come from highly ranked cities. In Figure. 2.2 we show the Top 20 cities from 1990 to 2009. Interestingly, we clearly see the decline and rise of cities along the years as well as the steady leadership of Boston and Berkeley. This behavior is clear in Figure. 2.3-B where we show the rank for cities in USA in 1990 and 2009. Meanwhile, the ranking of cities in European and Asian countries like France, Italy and Japan has increased significantly, as shown in both Figure. 2.2 and Figure. 2.3-A. In Figure. 2.3-C we focus on the geographical distribution of ranks for a selected set of European countries in 1990 and 2009.

3 Detecting and predicting seasonal flu using Twitter data

Fast response to seasonal influenza epidemics is critical to reduce the spread of the disease and economical loss. Traditional surveillance systems of influenza-like illness (ILI) usually experience 1-2 weeks delay between the time a patient is diagnosed and the case is aggregated into ILI reporting system [3]. Thanks to the fast developed Internet search query engines, nowadays there are various online open tools as real-time surrogates for clinically-based reporting of influenza-like-illness. A typical example is Google Flu Trends, which have been used applied to generate on-time detection and estimation of influenza epidemics [37, 30]. However, it is also reported that Google Flu Trends model suffers substantial flaws such as overestimating the 2012/2013 influenza epidemic, and cannot be an ideal substitute for local surveillance [61]. Meanwhile, GPS enabled mobile clients for microblogging platforms and social networking sites allow for the quantitative analysis of spatial and temporal information of social systems. For instance, Twittter, a popular microblogging platform, can provide a good resource for detecting real time events and has been applied in flu detection [27, 3, 52]. These existing works on flu detection, however, focus on fitting Twitter signals with the surveillance data and show a lack of considering epidemic dynamics. In this project, we consider both geolocalized tweets containing ILI related keywords as an indicator of seasonal flu outbreak and detailed spreading dynamics with a stochastic epidemic model GLEAM. We aim to predict ILI cases in the

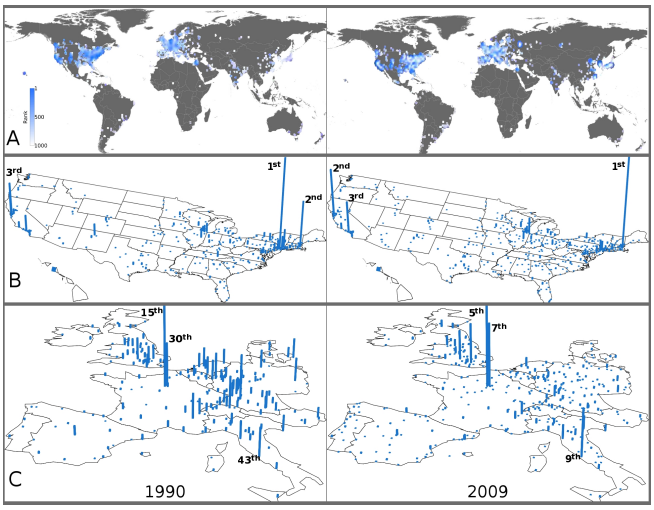


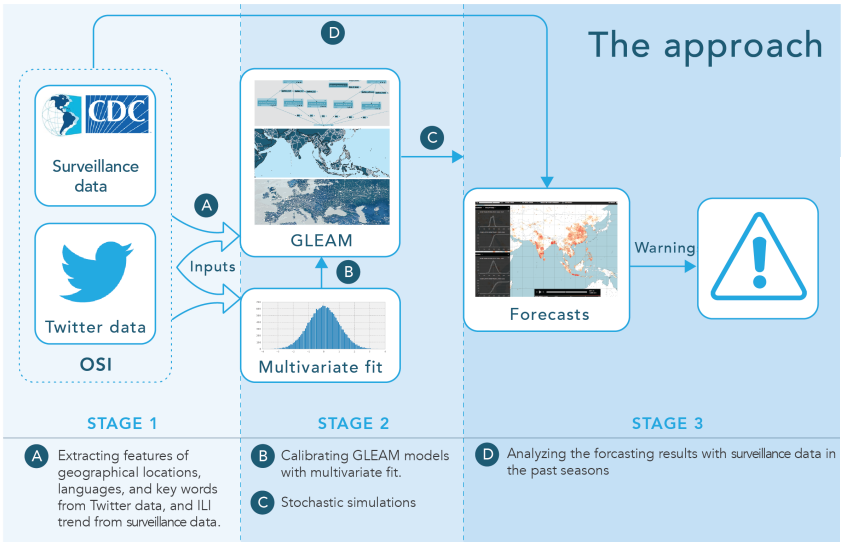
Figure 2.3: **Geospatial distribution of city ranks.** (A) The world map of city ranks in 1990 (left) and 2009 (right). The ranking of each city is represented by color from blue (high ranks) to white (low ranks). (B) The map of ranks for cities in the United States in 1990 (left) and 2009 (right). (C) The map of ranks for cities in the selected European countries in 1990 (left) and 2009 (right). In (B) and (C), each city is marked with a bar, and the height of each bar is inversely proportional to the ranking position. The Top 3 rank positions in each region are labeled for reference. Note that in (C) the height of bars is not scaled with the height in (B) for visibility.

coming season.

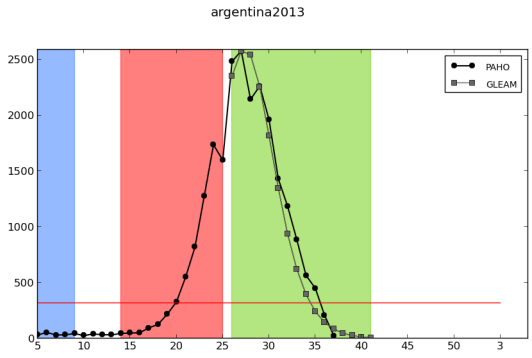
The global epidemic and mobility (GLEAM) model [9, 7, 8] is a data-driven global stochastic epidemic model based on the geographically structured metapopulation approach. The model integrates three different data layers. First, the population layer is composed of $3,362$ subpopulation areas in 220 countries in the world. Inside each subpopulation, the population is projected from "Gridded Population of the World" [20], which provides estimations of population for cells of 15×15 minutes of arc worldwide. Second, the transportation layer includes human mobility flows among subpopulations in both long range (airline traffic) and short range (commuting flows). The last one is the epidemic layer. Inside each subpopulation, the infection dynamics is simulated with a compartmental model. The transportation layer diffuses the exposed and infected individuals across subpopulations, thus allowing the stochastic simulation of the worldwide unfolding of the epidemic. The GLEAM model has been proved efficient to make short and long term predictions of the global unfolding of pandemic diseases [69] and is ready to be applied in warning generation in the unlikely case of a novel pandemic. In this project, we only consider such the disease spreading process within a specific country of interests.

Figure. 3.1 (a) illustrates the data flow from Twitter, surveillance data to GLEAM model and finally to the prediction. In stage I, we estimate the location and volume of initial infected seeds from geolocalized tweets and surveillance data. We fuse these initial seeds into GLEAM model in stage II and perform a large amount of stochastic simulations with different sets of parameters. Finally we choose the best set of parameter by fitting the simulation results with the known surveillance data (from Center of Disease Control (CDC) [1] for the United States and from Pan American Health Organization (PAHO) [2] for Latin American countries) and generate warnings for the coming weeks.

The first question is at which time we start to simulate epidemics. We defined a threshold on the number of cases above which the model is used to create warnings. The threshold is the baseline of normal activity and as shown in Figure. 3.1 (b) divides the timeline of each country into three time windows. In the first window (blue) the number of cases is below threshold. This window is used to estimate the initial seeds for the model and we call it seeding window. With initial seeds for a given week in this window, we can perform a large scale stochastic simulations using GLEAM, and prepare data for the next steps. In the second window (red) the number of cases crosses the threshold. This window is used to generate



(a) The illustration of the approach



(b) An example of fitting and predicting

Figure 3.1: (a) The schema of working flow to estimate initial seeds, calibrate GLEAM model and forecast. (b) An example of using simulated GLEAM results to predict ILI cases in the coming weeks.

refined warnings with GLEAM and the surveillance data, and is named as the fitting window. In the third window (green) the number of cases is above the threshold and we provide predicted number of cases from the refined warnings. The value of the threshold is selected in each country by considering the ILI activity the past seasons as done by the CDC.

3.1 Estimating initial seeds from Twitter signals

We consider geolocalized Twitter data originated in a specific country (e.g., the United States). In particular, we restrict our attention to those tweets for which the spatial information is in the format of geographical coordinates, either generated automatically through a GPS device (smartphone) or self-reported. With such geolocalized tweets, we can frame spatial and temporal information related with ILI keywords into stage I in Figure. 3.1 (a).

We consider a set of 5 weeks as the seeding period (blue region in Figure. 3.1 (b)). This window is fixed to be between 11-15 weeks before the threshold. After determining the seeding window, we can estimate initial seeds for each subpopulation area in the given country. We define $\omega_{l,w}^{C}$ to be the number of tweets matching ILI-related keyword l, in week w in a given country C. By fitting the profile of each keyword with the official surveillance data (e.g., CDC ILI report), we can evaluate R_l^2. This quantity tells us how correlated the two time series are. As described above, in GLEAM model, each country is divided into subpopulation areas centered in a major transportation hub. Using geolocalized tweets and the results from the fits at the country level, we can estimate the number of seeds for each starting week w, in each basin k, and country C as:

$$I_{k,w}^{C} = \left(\sum_{l} \omega_{l,w}^{C} R_l^2 \right) \alpha_k Y \tag{3}$$

The sum considers all the matches of ILI related keywords in the country C during a given week w. Each count is rescaled with the coefficient of determination R^2. This rescaling is necessary because it counts for the actual correlation between each keyword and the number of ILI cases. The coefficient α_k is the ratio of census population to the total number of Twitter users that are determined to live in subpopulation k. Y is a free parameter. Such the initial seeds will be feed into GLEAM model for stochastic simulations in stage II.

3.2 Calibrating GLEAM and fitting surveillance data

Inside each subpopulation, the disease dynamics is modeled with a Susceptible-Latent-Infectious-Recovered (SLIR) compartmental scheme, typical of ILI, where each individual has a discrete disease state assigned at each moment in time. Let β is the infection transmission rate and I_j/N_j is the density of infected individuals in the subpopulation j. Given the force of infection $\lambda_j = \beta I_j/N_j$, each person in the susceptible compartment (S_j) contracts the infection with probability $\lambda_j \Delta t$ and enters the latent compartment (L_j), where Δt is the time interval considered. Latent individuals exit the compartment with probability $\varepsilon \Delta t$, and transit to asymptomatic infectious compartment (I_j^a) with probability p_a or, with the complementary probability $1 - p_a$, become symptomatic infectious. Infectious persons with symptoms are further divided between those who can travel (I_j^t), probability p_t, and those who are travel-restricted (I_j^{nt}) with probability $1 - p_t$. All the infectious persons permanently recover with probability $\mu \Delta t$, entering the recovered compartment (R_j) in the next time step.

We then use GLEAM to perform a Latin Square Sampling in the parameter space $w \times Y \times r$, where w is the starting week, Y is the free parameter that rescales the initial number of cases, and r is the initial fraction of the immunized population. The simulation results of the models for each point in the sampled parameter space are compared with the real data in fitting window. This window, shown in red in Figure. 3.1 (b), is defined as the set of 12 weeks before the last data points of the surveillance data. Once the best set of parameter is selected, the weeks above the threshold in the predicting window (green) are considered to generate warnings.

When using the this approach to predict the ongoing flu season there is a complication that does not affect studies of historical data. Indeed, the scale of the number of cases provided by surveillance data and by GLEAM is extremely different. It is necessary to rescale them in order to fit and generate warnings. In historical data this is a simple task: GLEAM data is rescaled by a factor $x = \max(\text{GLEAM})/\max(\text{DATA})$. In the ongoing season the maximum (peak value) of the ILI cases in general may not be known, implying the impossibility of evaluating the rescaling factor x. If the surveillance data have reached the peak in the current season, we apply the same technique as for historical data. If the data have not reached the peak value, we consider the average peak values in the past seasons. Since the peak value can be different from season to season, we also consider the standard deviation from the peak values in the past seasons. Formally the rescaling factor $x = \max(\text{GLEAM})/(\langle H \rangle + k\sigma)$, where $\langle H \rangle$ is the average peak values in the past seasons for a given country, and we set $k = -1, 0, 1, 2$.

With rescaled GLEAM data for each point the sampled parameter space, we try to fit them with the surveillance data in the fitting window with different fitting methods. For each GLEAM candidate, we calculate the final fitting score s and select the one with the minimum value. s is defined as : $s^2 = \sum w_i (d_i - g_i)^2, i = 0, \ldots, M-1$ where d_i is the real data and g_i is the rescaled GLEAM data in the week i. M is the fitting window size and $i = 0$ means the first week in the fitting region. The simplest method is unweighted fitting, where $\forall i, w_i = 1$. This method treats all data points in the fitting region equally. However it is possible to argue that this is not the case. Indeed, getting better results in the last weeks, which are closer to the green window, might be more important than reproducing perfectly the first weeks. In order to include this consideration in the fits we tested different choices assigning weights to each week in the red window giving more importance to the latest weeks. We tried two different weight functions. For *exponential* weighted fitting $w_i = e^{(i-10)}$ and for *linear* weighted fitting $w_i = (i+1)/10$ if $i < 10$ else $w_i = 1$. These two methods give more importance on the last two points in the fitting windows. In Figure. 3.1 (b), we show an example of fitting and predicting for Argentina in season 2013. Suppose the last data point we have from the surveillance data is the week 25, and at the week 20 the surveillance data cross the threshold. The seeding window is week 5 to 9 and the fitting window is week 14 to 25. We rescale the GLEAM candidates with the peak values of this season, and find the GLEAM candidate that best fits the surveillance data in the red window among all points in the sampled parameter space and with two weighted fitting methods.

3.3 Future works

The method we propose above can provide reasonable results for some countries with testing historical data. However there are still some issues and some key factors that we have not taken into consideration and will be solved in the coming months.

I Currently we only use a list of ILI-related keywords to filter tweets. Some tweets containing these keywords might not indicate a single ILI case. For instance, if a news agency in a country tweets a news indicating an influenza outbreak in another country, and its followers can retweet the message. We now do not distinguish such tweets representing events from the tweets indicating single cases. To better understand the tweet contents and get well defined indicators of ILI, we plan to filter tweets indicating real ILI cases with techniques of natural language processing, instead of simply counting tweets containing keywords. We are trying to collect a random sample of tweets in the past seasons with at least one ILI-related keyword. With this sample, we are going to do human-annotating to get a training set of tweets representing single ILI case. With the trained dataset we plan to use the n-gram model classifying tweets.

II In the very early stage in a seasonal flu season, tweets containing ILI related keywords might be very random and may not reflect the real spatial distribution of initial seeds. Besides, we found that the final parameter set for the best fitted GLEAM simulation is not sensitive to the initial weeks. In order to get better estimation of initial conditions for GLEAM model, we plan to select the starting week of simulations to be two weeks before the surveillance data cross the baseline.

III We fixed the basic reproduction number $R_0 = 1.75$ in GLEAM model at the present time. R_0 for seasonal influenza might change from season to season, e.g., for the United States it varies from 0.9 to 2.1 [23]. We plan to perform more simulations and analysis with different values of R_0 and test the sensitivity of results to such model parameters.

IV Besides the ILI reporting at the country level, CDC also collects data for nine census regions. Weekly updates are also issued at regional granularities. Such high-resolution data provide an opportunity to test predictions at different geographical resolution. We are planning to apply the above approach to different regional datasets and aggregate to the country level. With this approach we will be able to test the geographical role that Twitter signals play at different spatial granularities.

4 Strength of weak ties on diffusion process on mobile communication networks

Beyond various studies on the dynamics of diffusion processes on complex networks [11], recent works on studying diffusion process on social networks have recourse to the traditional concepts in sociology, like weak ties [63, 73, 43]. According to [38], local bridges in a social network are responsible for the information transmission through the embedded communities and play a crucial role in information diffusion between otherwise disconnected communities since they are on the shortest path between many nodes [38, 39]. All of these bridges are weak ties, and the significance of weak ties is the local bridge creating more and shorter path [38]. Such definition of weak ties provides better understandings of how micro-interaction behavior translating into macro-patterns of a social system [38].

Thanks to the emergence of large datasets of human-to-human interactions, the strength of links on real large-scale social systems such as mobile communication networks and online social networks has been measured quantitatively with various definitions: overlap [62, 73], the number of calls, the maximum inter-event time between two individuals [44, 43], etc. These measurements consider either topological features of the static network or temporal patterns of events. These past studies confirm that weak ties help slowing down the spreading process [63] and the dynamical processes are not necessarily topologically efficient because of weight-topology correlation and burstiness of individuals [44]. Meanwhile it also reported that the strength of weak ties does not help the diffusion process of complex contagions [21, 73, 36]. However until now there is still no agreement on what is a "weak tie". In this study, we go back to the original Granovetter's definition of weak tie, and consider both the role each link plays on information diffusion process and its topological role on the network. We take

advantage of collections of human-to-human communication records in real life, and use an information propagation model to find out how important a link is in the dissemination process.

We take data of mobile communication records in one month from an unnamed country, and aggregate the records into a human-to-human communication network. In the following, we use the k-core ($k = 2$) of the largest connected component of the network to avoid isolated nodes and possible bias of calculating the strength of ties due to leaf nodes. The final network includes $1,926,787$ nodes and $3,269,634$ links. We simulate Susceptible-Infected process on this network by using the the number of mobile calls between each pair of individuals as the link weight. All nodes are susceptible initially except a randomly selected seed. Information is allowed to spread from one individual to another. In one single realization, the process continues until all nodes receive the information, and we record the spreading tree rooted from the initial seed. Since we allow a node to participate in more than one communication simultaneously (and in reality it also happens), the resulted graph is strictly a tree. We define a novel link property *importance*, which is calculated with the river basin algorithm [66]:

Algorithm 1: River-basin algorithm to calculate the *importance* of links from SI spreading tree.

INPUT : the spreading tree $T = (E, V)$
OUTPUT: the counters indicating *importance* for links
for $e = (i, j) \in E$ **do**
  initialize a counter $C_{i,j} = 0$
while $T \neq \emptyset$ **do**
  Leaves $= \emptyset$
  for $v \in V$ **do**
    if $\text{degree}_v == 1$ **then**
      for $(u, v) \in E$ **do**
        $C_{u,v} \longleftarrow C_{u,v} + 1$
        for $(r, u) \in E$ **do**
          $C_{r,u} \longleftarrow C_{r,u} + C_{u,v}$
      Leaves $\longleftarrow$ Leaves $\cup \{v\}$
  $V \longleftarrow V -$ Leaves
  $E \longleftarrow E - \{(k, l) | \forall l \in \text{Leaves} \wedge (k, l) \in E\}$
  $T \longleftarrow (E, V)$

Now the count for each link, *importance*, equals to the number of nodes which received information via this link. This link property sticks to Granovetter's definition: links with high values of such the count tend to serve as the shortest path for information propagation between nodes and play the role of local bridges in the SI process, and therefore are weak ties according to Granovetter's descriptions [38, 39]. We perform multiple realizations (100) of such process and get the average *importance* for each link.

In this study, we focus on this new definition of strength of ties *importance* and investigate the effects of weak ties on the diffusion process. Meanwhile we also consider the other definitions of the strength of links: the number of calls between two individuals in the aggregating time window W_{Ncall} and the topological feature overlap $O_{ij} = n_{ij}/((k_i - 1) + (k_j - 1) - n_{ij})$ [63] where n_{ij} is the number of common neighbors between i and j. In Figure. 4.1, we first report the correlations between these three types of definition of link strength. Figure. 4.1(a) and (b) show the behavior of the average overlap as a function of the cumulative link strength defined by W_{Ncall} and *importance* respectively. The cumulative link strength is defined by $P_c(x) = \int_{-\infty}^{x} p(w)dw$, where $p(w)$ is the probability density function for the link strength (W_{Ncall} or I) [62]. The positive correlation between *overlap* and the number of calls W_{Ncall} is also reported in [62], that verified the weak ties hypothesis of Granovetter, "the strength of a tie is a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" [38]. The negative correlation between

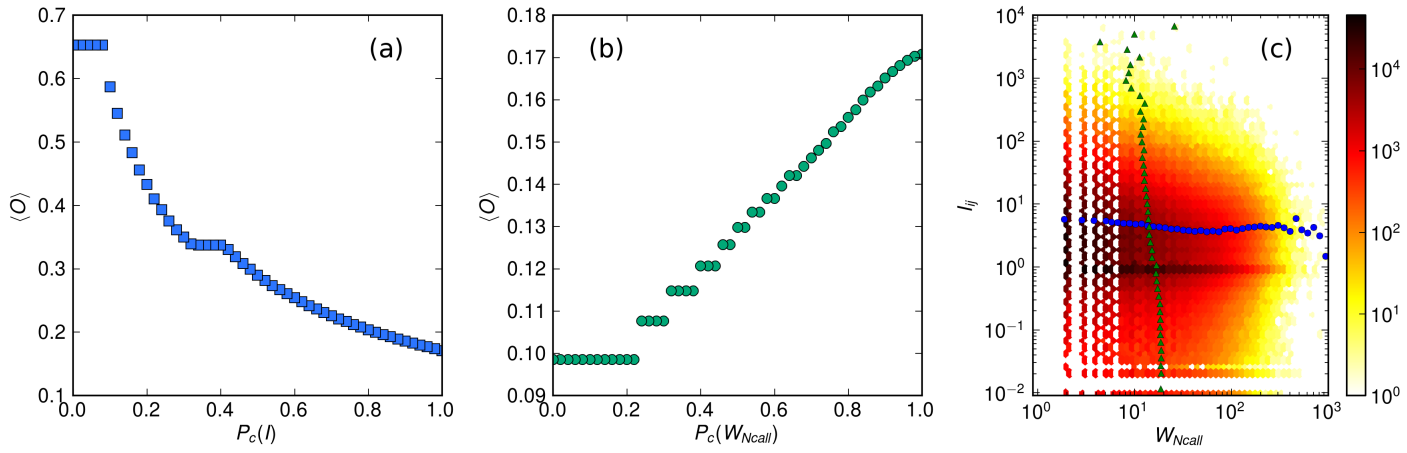


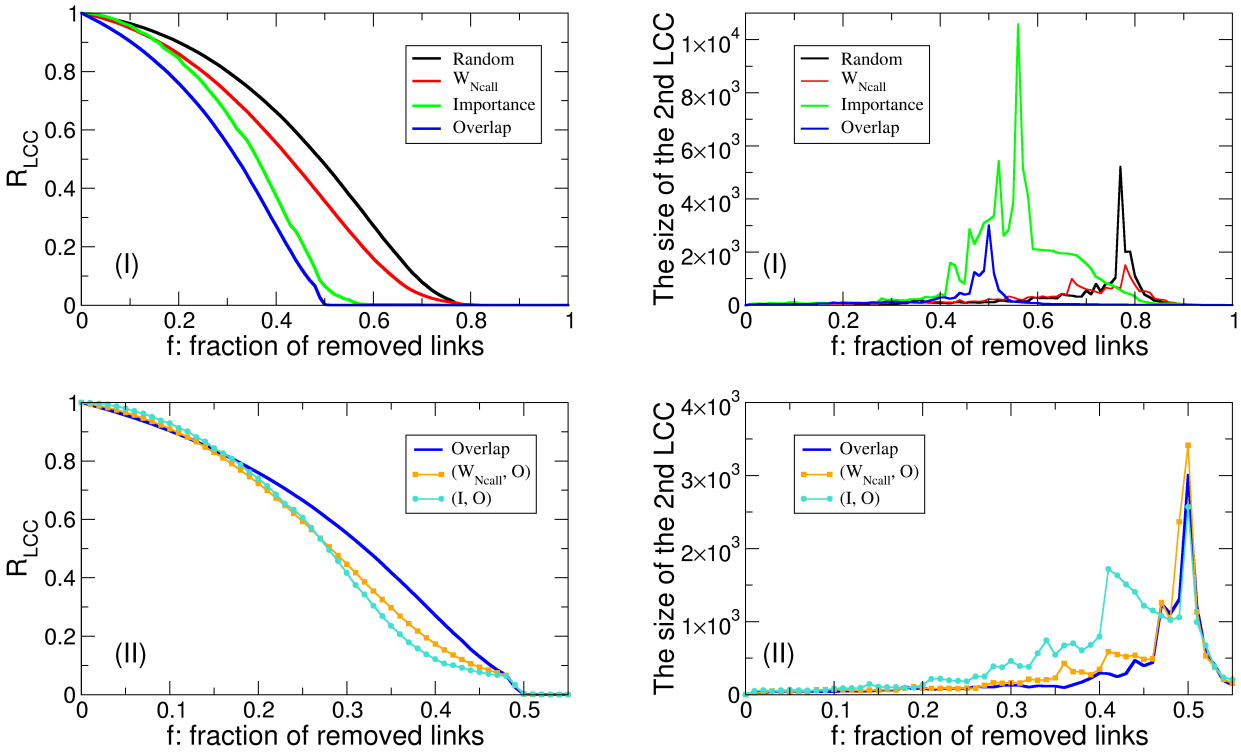
Figure 4.1: Correlations between various definitions of weak ties. (a) The average overlap as a function of cumulative link *importance*; (b) the average overlap as a function of cumulative link weights defined by the number of calls between nodes W_{Ncall}; (c) correlation between the weight W_{Ncall} of the links (x-axis) and the *importance* I of the same links (y-axis), where the blue dots represent the average *importance* as a function of W_{Ncall}, and the green triangles refer to the average W_{Ncall} as a function of *importance* I. Although overlap is positively correlated with W_{Ncall} and negatively correlated with I, W_{Ncall} and *importance* I do not show any correlation.

overlap and *importance* is also straightforward: the end nodes of links with low values of overlap have small fraction of common neighbors, and these links tend to connect different cliques and serve as local bridges in the diffusion process. In Figure. 4.1(c), we show the correlation between the weight W_{Ncall} of the links and the *importance* I of the same links. Although overlap is positively correlated with W_{Ncall} and negatively correlated with I, W_{Ncall} and *importance* I do not show any correlation. Although the calculation of *importance* is based on dynamical processes on the network with weight W_{Ncall}, it is independent of the weight and only reflects the topological feature of links and the role the link plays in the diffusion process. In the following section, we investigate the structural roles of weak ties and the effect of controlling weak ties on the diffusion control under different definitions of strength.

4.1 Structural role of weak ties

Removing weak ties results in a phase transition-driven collapse of the network [63]. One possible way to investigate the Granovetter feature of links in the network is to perform a link removal process and compare critical percolation behaviors to a corresponding random link removal process. For each definition of weak tie $W_{\mathrm{Ncall}_{ij}}$, O_{ij} and I_{ij}, we order the links according to the weakness, and for combinations of link properties $((W_{\mathrm{Ncall}_{ij}}, O_{ij}), (I_{ij}^{-1}, O_{ij}))$ we first order links from the lowest to the highest value of overlap and for the links with the same value of overlap, we order them based on the weakness. Removing process is controlled by the ratio of removed links f. $f = 0$ refers to the initial connected network and $f = 1$ represents a set of isolated nodes.

To evaluate the impacts of removing different types of weak links, we first study the relative size of the largest connected component (R_{LCC}) as a function of the ratio of links to be removed f in Figure. 4.2 (a). We find that removing links by ordering overlap leads the network collapse near $f_c^o = 0.5$. Removing links by the combinations of link properties ($(W_{\mathrm{Ncall}_{ij}}, O_{ij})$ and (I_{ij}^{-1}, O_{ij})) results in the same critical point as the removal strategy of overlap. In Figure. 4.2 (b), we also report the size of the second largest connected component as a function of the ratio f. The size of the second largest connected component reaches the maximum value at the percolation threshold [68]. Removing links by ordering O_{ij} and by ordering the combinations of link properties ($(W_{\mathrm{Ncall}_{ij}}, O_{ij})$ and (I_{ij}^{-1}, O_{ij})) lead the size of the second largest connected component reaching the maximum value at $f \approx 0.5$. Removing links by ordering other single definitions of weak ties makes the network collapse



(a) The relative size of the largest connected component

(b) The size of the second largest connected component

Figure 4.2: Percolation study: Global quantities of the network as a function of the ratio of links to be removed f.

at $f_c > 0.5$. If links are removed according to the weakness defined by importance, $f_c^I \approx 0.56$. Meanwhile, removing links that ordered by the number of calls disintegrates the network at $f_c^{W_{\rm Ncall}} \approx 0.75$, close to the critical value for random removing links.

In Figure. 4.2 (a-II) we observe that the network collapses faster if links are ordered by the combinations (I_{ij}^{-1}, O_{ij}) and $(W_{{\rm Ncall}_{ij}}, O_{ij})$ than removing links ordered by overlap alone when the fraction of removed links is between 0.2 to 0.48. From the perspective of percolation, the link property overlap determines the critical point of the phase transition. A link with $O_{ij} = 0$ indicating the nodes it connects has no common neighbors, and connects different social cliques or local communities. In the sample of mobile call network we investigate, there are around 48% of links with $O_{ij} = 0$. Removing all of these links disintegrates the network into a set of cliques. Removing links by ordering the combinations of weakness properties of link does not change the percolation threshold but speeds up the network collapsing, i.e., removing a fraction of links $f < 0.48$ results in the smaller size of the largest connected component.

4.2 Diffusion role of weak ties

In information spreading process, such weak ties with $O_{ij} = 0$ enhance the trapping effect inside local clusters, and slow down diffusion process between clusters [62]. Controlling these weak ties could further impede the spreading. For instance, in the case of rumor spreading in a communication network or infectious disease spreading in a contagion network, one could expect to weaken the strength of a small fraction of links to impede the propagations. Further identifying weak ties

among these zero overlap links could reduce the amount of links to be controlled. In this section, we turn into investigating the effects of controlling weak links under various definitions on diffusion processes to study how much we can control the contagion by decreasing the transmission probabilities on weak ties.

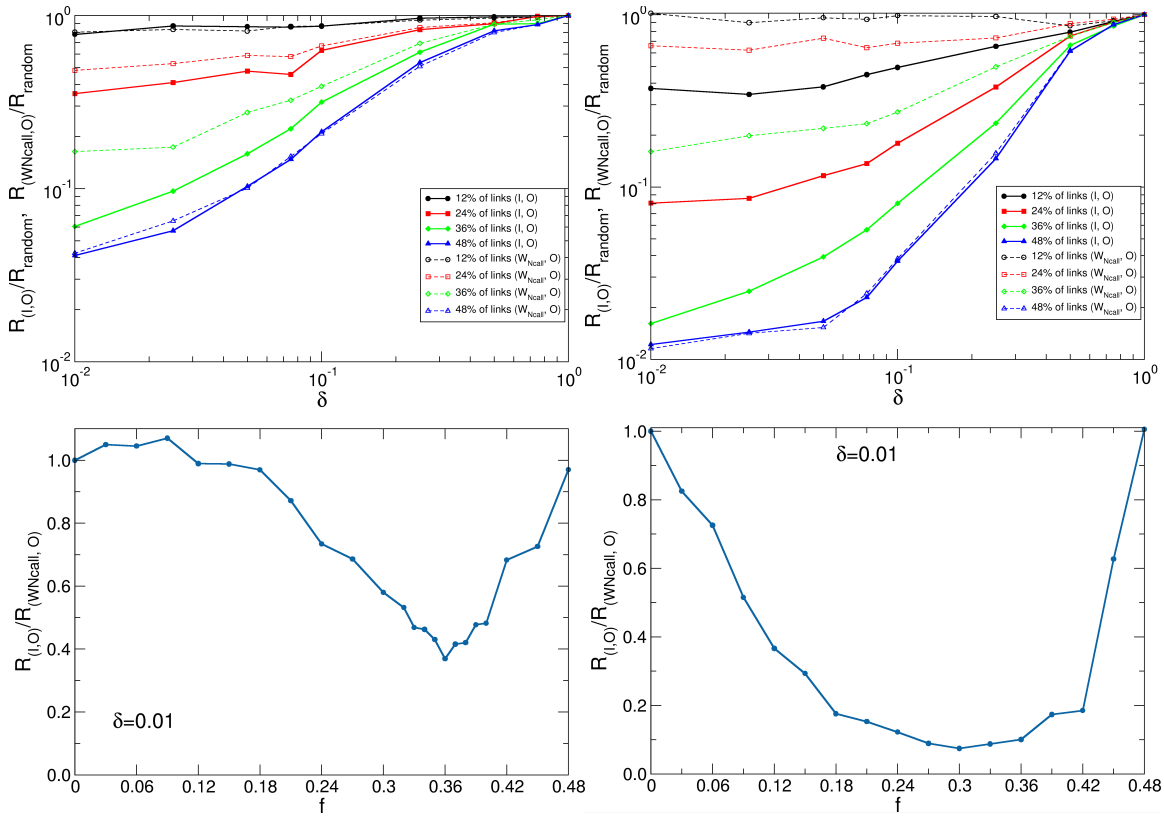
From the results above, removing links by ordering their property combinations $(W_{\text{Ncall}_{ij}}, O_{ij})$ or (I_{ij}^{-1}, O_{ij}) results in network collapsing faster than other definitions of weakness when the fraction of links removed is between 0.2 and 0.48. To impede the diffusion process of rumor or infectious disease, one could decrease the probabilities of contagion between two nodes between the weak link by weakening their connection weight. SIR model is well known for studying dynamics of epidemics [4] and rumor spreading [59]. In the following we apply SIR model on the communication network and control weak ties defined by the combinations of link properties by weakening the weights of links. Consider the SIR system $S + I \xrightarrow{\beta} 2I$, $I \xrightarrow{\mu} R$. The final size of epidemic R refers to the total recovered population size. We fixed the infection rate $\beta = 0.25$ and the recovery rate $\mu = 0.1$. The recovery rate μ is constant for each individual in the network. However, the probability of an individual i to be infected by contacting j is $p_{ij} \propto w_{ij}\beta$. On the unweighted network $w_{ij} = 1$, and on the weighted network w_{ij} is defined based on the number of calls between the two nodes (W_{Ncall}).

We first focus on the unweighted network. Similarly to the percolation study, for a given percentage of links f, we order the first f links based on the combination of link properties (I_{ij}^{-1}, O_{ij}), and we scale the link weights w_{ij} from 1 to $\delta, (0 < \delta \leq 1)$. As a control experiment, we randomly choose the same percentage of links f, and scale the link in the same way. We compare the final size of epidemic in these two cases by measuring $R_{(I^{-1},O)}/R_{\text{random}}$ as a function of the scaling factors δ for different values of f. If controlling weak ties performs as the same as the random links, this ratio is 1.0. If it is more efficient, the ratio should be less than 1.0. In Figure. 4.3 (a) top panel (solid lines), we show such the comparison results. When $\delta \leq 0.1$, weakening the weights on weak ties reduces at least 50% of the final size of epidemic than weakening the weights on random links.

In the same figure, instead of (I_{ij}^{-1}, O_{ij}), we control weak ties that are defined by the $(W_{\text{Ncall}_{ij}}, O_{ij})$ and measure the same quantities for different values of f (dashed lines in Figure. 4.3 (a) top panel). Since there are around 48% links with $O_{ij} = 0$ the effects of controlling weak ties of these two definitions are similar. When $f = 36\%$, controlling weak ties defined by (I_{ij}^{-1}, O_{ij}) reduces the final size of epidemic more than 60% with the scaling factor $\delta = 0.01$. When $f = 12\%$, the effects of controlling weak ties by these two definitions are also similar.
To further compare the effects of controlling ties of these two definitions ($(W_{\text{Ncall}_{ij}}, O_{ij})$ and (I_{ij}^{-1}, O_{ij})) on spreading dynamics, we report $R_{(I^{-1},O)}/R_{(W_{\text{Ncall}_{ij}},O_{ij})}$ as a function of the fraction of controlled links f for a given $\delta = 0.01$ in Figure. 4.3 (a) bottom panel. When $0.24 < f < 0.48$, one can observe controlling weak ties defined by (I_{ij}^{-1}, O_{ij}) results significantly less infected population than by $(W_{\text{Ncall}_{ij}}, O_{ij})$. This result is consistent with the observation in percolation study, that when f is between 0.25 and 0.45, the largest connected component is smaller if removing links are ordered by (I_{ij}^{-1}, O_{ij}). Controlling weak ties defined by (I_{ij}^{-1}, O_{ij}) limits the infected population inside a smaller community.

We also perform the same experiments on the weighted mobile communication network. The weight between node i and node j is defined as $w_{ij} = W_{ij}^{\text{NCall}}/\langle W^{\text{NCall}}\rangle$ if $W_{ij}^{\text{NCall}} \leq \langle W^{\text{NCall}}\rangle$, and $w_{ij} = 1$ if $W_{ij}^{\text{NCall}} > \langle W^{\text{NCall}}\rangle$, where W_{ij}^{NCall} is the number of calls between node i and node j and $\langle W^{\text{NCall}}\rangle$ is the average number of calls among all pairs of individuals in the network. Figure. 4.3 (b) top panel shows the ratio $R_{(I^{-1},O)}/R_{\text{random}}$ as a function of the scaling factor δ for the SIR process on the weighted network. The results are similar to the unweighted case. For example, at $\delta = 0.01$, controlling 12% of weak ties defined by $R_{(I^{-1},O)}$ leads to 60% less infected population than randomly controlling the same fraction of links. When $f = 24\%$ and above the final size of epidemic reduces more than 90% than controlling with the random strategy for $\delta < 0.1$. Similar to the unweighted case, we also report $R_{(I^{-1},O)}/R_{(W_{\text{Ncall}_{ij}},O_{ij})}$ as a function of the fraction of controlled links f for a given $\delta = 0.01$ in the bottom panel in Figure. 4.3 (b). When $0.18 < f < 0.48$, controlling weak ties defined by (I_{ij}^{-1}, O_{ij}) results at least 80% less infected population than controlling weak ties defined by $(W_{\text{Ncall}_{ij}}, O_{ij})$. Although the quantitative results vary, the qualitative behaviors in the experiments on both unweighted and weighted networks are

similar. The definition of weak ties is independent from the definition of weights, therefore we could claim that the results we obtained from controlling weak ties on the weighted network is not a direct result from the arbitrary definition of weights.



(a) unweighted, $\beta = 0.25, \mu = 0.1$ (b) weighted $\beta = 0.25, \mu = 0.1$

Figure 4.3: Diffusion control: scaling weak ties to control contagion process. For the unweighted network (a) and the weighted network (b), we scale down the link weights with the strategy of controlling weak ties defined by (I_{ij}^{-1}, O_{ij}) and $(W_{\mathrm{Ncall}_{ij}}, O_{ij})$ and with the strategy of randomly selecting link (top panel). We show the ratio of the final size of epidemic by scaling down link weights of weak ties and by scaling down random links. In the bottom panel of (a) and (b) we also report the ratio of the final size of epidemic by scaling down weak ties defined by (I_{ij}^{-1}, O_{ij}) and $(W_{\mathrm{Ncall}_{ij}}, O_{ij})$ for unweighted and weighted network respectively.

4.3 Discussion

In this project, we take advantage of a large scale dataset of mobile communication, and define a novel quantitative link property *importance* to measure the strength of links in a human-to-human communication network. The new definition of the strength sticks to the original Granovetter's qualitative definition of weak tie, and considers both the role each link plays on information diffusion process and its topological role on the network. We investigate the structural roles of weak ties and the effect of controlling weak ties on the diffusion control under different definitions of strength. We found the combination of *importance* and *overlap* (I_{ij}^{-1}, O_{ij}) is the best indicator of weak ties. Removing links ordered by the combination of *importance* and the overlap results in the same critical point as removing links ordered by the single definition of overlap. However, the network collapsing faster with the combination definition of weak ties when the fraction of removed links is between 0.2 and 0.48. We further investigate the role these links play on the diffusion process. We showed that controlling the weak ties defined by the combination of (I_{ij}^{-1}, O_{ij}) can impede diffusion process more efficiently than controlling the weak ties defined by other definitions. The links with high importance and zero overlap play the crucial topological role in connecting small communities, and propagating information into these communities. This definition of link property help us

gaining a better understanding of the weak ties in human-to-human communication networks.

5 Phase transition of rumor spreading process on complex networks

In order to study information diffusion on complex contact networks we have considered the rumor model by Maki and Thompson [53]. The model defines three exclusive states or compartments for each individual: ignorant (I), spreader (S) and stifler (R). Ignorants are those individuals who have not heard the rumor yet. While both the spreaders and stiflers have heard the rumor, only the spreaders wish to spread it. The number of ignorants $I(t)$, spreaders $S(t)$ and stiflers $R(t)$ at time t hence constitute the state variables of the system. An ignorant hears the rumor from a spreader neighbor and becomes a spreader herself at rate λ. Each spreader becomes a stifler, hence, stops spreading the rumor either by a spreader or a stifler contact at rate α. This set of rules defines a reaction or diffusion process in which $I + S \xrightarrow{\lambda} 2S$, $S + S \xrightarrow{\alpha} R + S$ and $S + R \xrightarrow{\alpha} 2R$. In the absence of demographic changes the total population size N is conserved and $N = I(t) + S(t) + R(t)$. In the discussion below $i(t)$, $s(t)$ and $r(t)$ refer to the prevalence of ignorant, spreader and stifler individuals in the population, respectively.

While the above dynamics is reminiscent of the well-known Susceptible-Infected-Recovered epidemic model [46], it has a subtle difference. Unlike the epidemic model, the rumor model does not possess a spreading threshold when the population is fully mixed, i.e., each individual is capable of contacting any other individual in the population. This means that whenever the spreading rate λ is nonzero, the rumor spreads to a finite fraction of the population in thermodynamic limit $N \to \infty$.

The original study by Moreno et al. [57] focusing on contact networks, where each individual has a fixed set of neighbors, has also concluded that the same is valid in homogeneous contact networks. However, the analytical approach taken in the study [57] is based on a Mean-Field assumption unable to tackle the quenched nature of connections. The availability of large-scale datasets on communication networks in recent years [62, 50, 48, 18, 49], however, presents new opportunities for the study of information dissemination in social networks and the utility of rumor models [53, 28]. In this project, we report numerical and analytical evidences that the parameters of the rumor contagion process and the architecture of connections determine whether a new rumor spreads globally in social networks.

5.1 Results from Monte Carlo simulations

To easily compare simulation results and analytical calculations, we consider uncorrelated random networks [19] in which each of N nodes has the same number k of connections. We generate at least 10 random networks with $k = 7$ for each population size N varying between 10^3 and 10^6. We fix the annihilation rate α at 0.5 and vary the spreading rate λ. For each set of parameters we perform at least 10^3 realizations of the rumor contagion process on each synthetic network. Each dynamical realization starts with a fully ignorant population except for one randomly chosen node turning into a spreader and runs until no spreader is left in the network. During the time step between t and $t + 1$, a spreader at time t contacts or calls all her immediate neighbors one by one. If the recipient is an ignorant at time t, the ignorant neighbor becomes a spreader at time $t + 1$ with probability λ. If the recipient is a spreader or stifler instead, the caller becomes a stifler at time $t + 1$ with probability α. In order to let the model be analytically tractable we do not impose any memory in the communications during a time step. Hence if a spreader calls another spreader, the callee turns into a stifler at time $t + 1$ with probability α, independent of the status change of the caller.

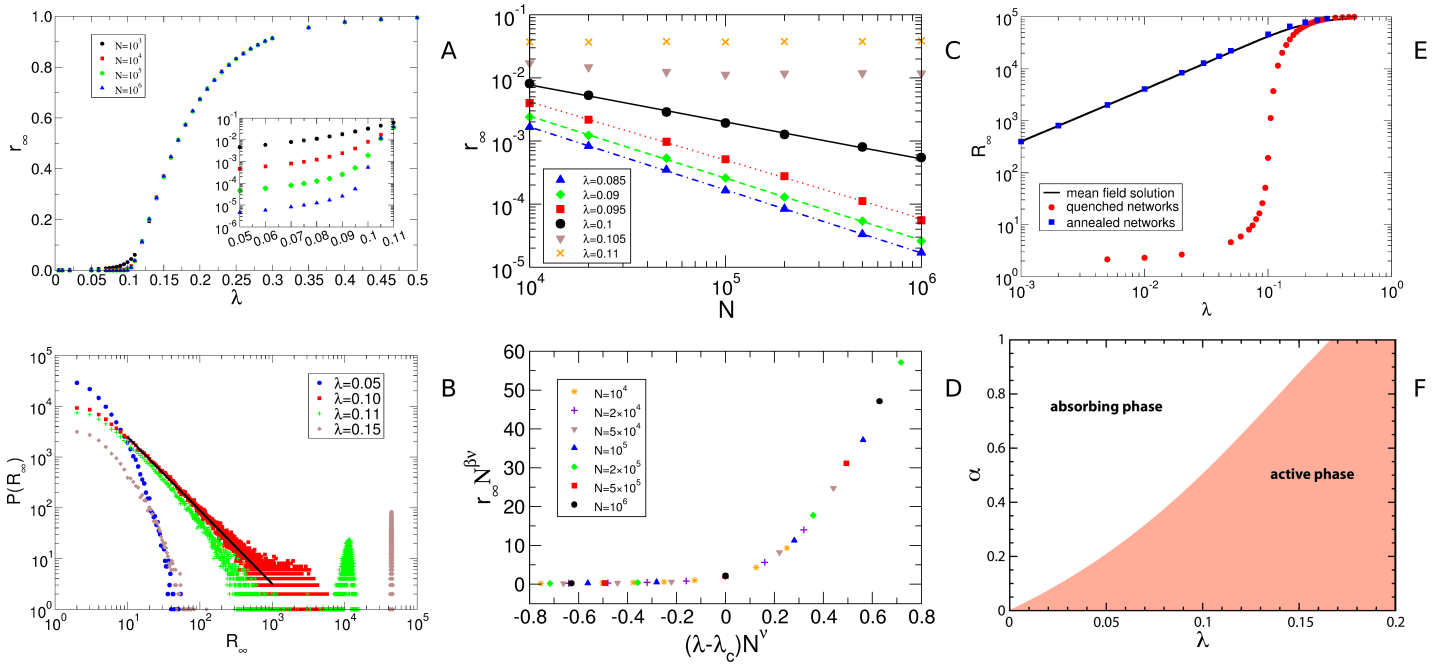


Figure 5.1: Monte Carlo simulation results (the results are obtained by averaging over 10×10^3 realizations, corresponding to 10 network realizations and 10^3 rumor contagion processes on each one of them) and numerical solution to analytic results. (A) The final prevalence of the rumor r_∞ as a function of the spreading rate λ for different population sizes $N = 10^3$ (black dots), $N = 10^4$ (red squares), $N = 10^5$ (green diamond) and $N = 10^6$ (blue triangles). The annihilation rate and the node degrees are fixed at $\alpha = 0.5$ and $k = 7$, respectively. The inset is a zoom of the transition region. (B) Probability distribution of the final rumor size R_∞ for $N = 10^5$, $\alpha = 0.5$, and $k = 7$. Different symbols correspond to different values of the spreading rate: $\lambda = 0.05$ (blue circles), $\lambda = 0.1$ (red squares), $\lambda = 0.11$ (green pluses) and $\lambda = 0.15$ (brown diamonds). The solid line corresponds to the power-law function with exponent -1.44 plotted for comparison with the distribution for $\lambda = 0.1$. (C, D) Finite-size scaling and data collapse. (C) The final rumor prevalence r_∞ as a function of population size N for different values of λ. The dashed-dotted, dashed, dotted and solid lines correspond to the power-law fits with the exponent -1, -0.99, -0.92 and -0.59, respectively. (D) The best data collapse of the final rumor prevalence r_∞ obtained at $\beta\nu = 0.59$, $\nu = 0.35$ and $\lambda_c = 0.1$. (E) The final prevalence of rumor r_∞ as a function of the spreading rate λ for the population size $N = 10^5$ on the quenched random network with (red dots) and on the corresponding annealed networks (blue squares). The solid line corresponds to the Mean-Field solution obtained from Eqs. 5–7 numerically. (F) Phase space of the rumor model on the λ–α plane for $k = 7$. The line separating the active and absorbing phases corresponds to $\bar{n}_\infty = 1$ and is obtained by solving Eqs. 10 and 13 numerically.

In Figure. 5.1 (A) we show the final prevalence of rumor r_∞ as a function of the spreading rate λ for different population sizes. We observe that when λ is smaller than about 0.1, the final prevalence of rumor r_∞ depends on the population size N. In contrast, when λ is larger, r_∞ is independent of N. The probability distribution of the final rumor size R_∞ for different values of λ, illustrated in Figure. 5.1 (B), shows the power-law distribution of the final rumor size for $\lambda = 0.1$. When $\lambda = 0.05$, the distribution decays exponentially. For $\lambda = 0.11$ and 0.15, the distributions are bimodal in that most of the realizations yields global rumor spread. All of these features are fingerprints of critical phenomena and imply the existence of non-equilibrium phase transitions in rumor model as the parameter space is explored. In the region of $\lambda \ll 0.1$ the spreading rate is much lower than the annihilation rate, and an early spreader loses her interest in the rumor much quicker than she can spread it. Thus the rumor is heard by a small number of early spreaders only, and the final size of rumor R_∞ does not scale with population size, i.e., the final prevalence of rumor $r_\infty \propto N^{-1}$. In the second region of $\lambda \gg 0.1$, the rumor can reach a large number of indivuduals because of the relatively large spreading rate combined with the number of connections. The final size of rumor in this region is proportional to the population size, i.e., r_∞ is independent of N.
Right above the critical point λ_c, the order parameter r_∞ obeys the scaling law $r_\infty \propto (\lambda - \lambda_c)^\beta$ [54]. In order to determine λ_c and β, we perform finite-size scaling analysis [54, 60]. In Figure. 5.1 (C), we show the final prevalence of rumor r_∞ as a function of population sizes N for λ values in the vicinity of 0.1. For $\lambda = 0.085$, 0.09 and 0.095, we find approximately $r_\infty \propto N^{-1}$. When $\lambda = 0.105$ and 0.11, the rumor reaches a finite fraction of the population that is independent of the

population size. These findings imply that the system is in a subcritical regime for $\lambda < 0.1$ and is supercritical when $\lambda > 0.1$. At $\lambda = 0.1$, $r_\infty \propto N^{-\rho}$ where $\rho = 0.59 \pm 0.02$, and we get the estimation of $\lambda_c = 0.1 \pm 0.005$. Following the finite-size scaling method [60]

$$r_\infty N^{\beta\nu} = F((\lambda - \lambda_c)N^\nu) \,, \tag{4}$$

where $F(\bullet)$ is the scaling function, and $r_\infty = F(0)N^{-\beta\nu}$ at $\lambda = \lambda_c$. As discussed above, $\rho = \beta\nu = 0.59 \pm 0.02$. In Figure. 5.1 (D) we display the best data collapse obtained for $\nu = 0.35 \pm 0.01$ in the vicinity of $\lambda_c = 0.1$, yielding $\beta = 1.68 \pm 0.01$.

The simulations show that the rumor propagation gets confined to a small neighborhood of early spreaders if the spreading rate is much smaller than the annihilation rate. This result is in contradiction with the previous studies [57] based on a Mean-Field approximation. As already stated earlier, the Mean-Field approximation considers an ensemble of annealed random networks [29, 14]. This means that the links are re-established while the degree distribution is kept invariant. In such an ensemble, every node has the potential of contacting every other node in the population during the course of a dynamical process running on top. The Mean-Field rate equations describing the prevalences of compartments over time are then roughly given by

$$\begin{aligned}
i'(t) &= -i(t)[1 - (1 - \lambda s(t))^k] \,, & (5)\\
s'(t) &= i(t)[1 - (1 - \lambda s(t))^k] - s(t)[1 - [1 - \alpha(s(t) + r(t))]^k] \,, & (6)\\
r'(t) &= s(t)[1 - [1 - \alpha(s(t) + r(t))]^k] \,, & (7)
\end{aligned}$$

with the constraint $i(t) + s(t) + r(t) = 1$. When λ and α are sufficiently small, the above equations are equivalent to the Mean-Field equations in [57].

In Figure. 5.1 (E) we report the simulation results of rumor propagation dynamics on annealed random networks of size $N = 10^5$. The simulations on annealed networks agree with the numerical solutions of the Mean-Field Eqs. 5–7, both of which are strikingly different from the simulations on quenched networks. The simulations on annealed networks have been performed by reshuffling all the edges at the end of each time step of the rumor dynamics. The reshuffling process enables early spreaders to keep their interests in the rumor and spread it to different neighborhoods until a sufficiently large number of individuals hear the rumor.

5.2 Analytic results

By definition of the rumor model, the initial spreader has to pass the rumor to a neighbor in order to enable the annihilation process. However, as the above simulation results imply, the outreach of the rumor on quenched contact networks is determined by the network structure and the spreading and annihilation rates. We may tackle the problem analytically by considering the immediate social network of an early spreader, hence, the ego network. Consider the moment at which the ego has x ignorant neighbors only among her k neighbors in total, hence, $k - x$ of her neighbors are either spreaders or stiflers. During the next time step the ego spreads the rumor to $0 \leq n \leq x$ ignorant neighbors and either loses her interest in the rumor or keeps her status as spreader. The probabilities of encountering these two events p_k and q_k, respectively, are given by

$$\begin{aligned}
p_k(n|x) &= \binom{x}{n} \lambda^n (1 - \lambda)^{x-n} (1 - \alpha)^{k-x} \,, & (8)\\
q_k(n|x) &= \binom{x}{n} \lambda^n (1 - \lambda)^{x-n} \left[1 - (1 - \alpha)^{k-x}\right] \,. & (9)
\end{aligned}$$

In order to answer the question concerning the global rumor spread, we need to consider the whole period in which the ego is a spreader. If an early spreader can spread the rumor to at least one ignorant neighbor on average, then the rumor may

have a global outreach.

Suppose that only the x_0 of k neighbors are ignorant at the moment when the ego's status turns into a spreader, and denote the probability that the ego spreads the rumor to n_∞ neighbors during her spreader period by $P_k(n_\infty|x_0)$. The average number of newly generated spreaders by an early spreader, i.e., $x_0 = k-1$, is then

$$\bar{n}_\infty = \sum_{n=0}^{k-1} n P_k(n|k-1) \; . \tag{10}$$

If $\bar{n}_\infty \geq 1$, then the rumor stays alive in the early stages of the contagion dynamics and be propagated globally. The rumor is going to be confined to a small neighborhood if $\bar{n}_\infty < 1$. It is worth remarking that the threshold parameter $\bar{n}_\infty$ is equivalent to the so-called basic reproduction number in epidemic processes [45]. The basic reproduction number is the average total number of secondary infections generated by an infectious person in the early stages of the epidemics. Equivalently here, $\bar{n}_\infty$ is the average total number of spreaders generated by a spreader during the early stages of the rumor dynamics.

While finding a closed form of the probability $P_k(n|x)$ is not a piece of cake, we can write down an expression that can then be computed numerically. Let us write down the expressions for $n = 1$ in order to demonstrate the logic here. The probability of spreading the rumor to one neighbor only is

$$P_k(1|k-1) \;\; = \;\; \sum_{t_1=0}^{\infty} [p_k(0|k-1)]^{t_1} \left[q_k(1|k-1) + p_k(1|k-1) \sum_{t_2=0}^{\infty} [p_k(0|k-2)]^{t_2} q_k(0|k-2) \right] \; , \tag{11}$$

where $t_1 + 1$ counts the number of time steps from the moment when the ego becomes spreader till the end of the time step during which the rumor is propagated to one neighbor successfully. The ego may become a stifler during the same time step. If the ego is still a spreader, t_2 counts the number of time steps proceeding the generation of the new spreader until the annihilation of the ego. We may easily perform the summations over the dummy variables t_1 and t_2 as these are just sums of geometric series, yielding

$$P_k(1|k-1) \;\; = \;\; \frac{1}{1-p_k(0|k-1)} \left[q_k(1|k-1) + p_k(1|k-1) \frac{q_k(0|k-2)}{1-p_k(0|k-2)} \right] \; . \tag{12}$$

We follow the same logic above in order to express the probability of spreading the rumor to an arbitrary number n of neighbors in total. The ego may generate all the n spreaders at once and become a stifler herself or survive for an additional period. The ego may spread the rumor to $n_1 > 0$ and $n - n_1 > 0$ ignorant neighbors sequentially. She may change her status during the last generation or survive a bit longer before losing her interest in the rumor. The ego can generate $n_1 > 0$, $n_2 > 0$ and $n - n_1 - n_2 > 0$ new spreaders sequentially. During the time step of the generation of the last group with size $n - n_1 - n_2$, she may become a stifler or keeps her interest in the rumor for an additional period. This continues until the last configuration in which each new spreader is generated sequentially. One may consider this sequence of events as the configurations of having n balls grouped into $1 \leq b \leq n$ boxes, each of which containing at least one ball, and a 'stop' sign aligned on the timeline. The 'stop' sign can be located only at/after the position of the last box on the timeline. Each ball corresponds to an ignorant neighbor who eventually hears the rumor from the ego, and each box to a time step during which the ego spreads the rumor to a certain number of ignorant neighbors. The stop sign is the event of successful annihilation. Considering all the configurations, the probability that the ego generates n new spreaders only during her spreader period is

given by Eq. 13:

$$
\begin{aligned}
P_k(n|k-1) &= \frac{1}{1-p_k(0|k-1)}\left[q_k(n|k-1)+(1-\delta_{n,0})p_k(n|k-1)\frac{q_k(0|k-1-n)}{1-p_k(0|k-1-n)}\right] \\
&+ (1-\delta_{n,0})(1-\delta_{n,1})\sum_{n_1=1}^{n-1}\frac{p_k(n_1|k-1)}{1-p_k(0|k-1)}\times\frac{1}{1-p_k(0|k-1-n_1)}\Bigg[q_k(n-n_1|k-1-n_1) \\
&+ p_k(n-n_1|k-1-n_1)\frac{q_k(0|k-1-n)}{1-p_k(0|k-1-n)}\Bigg] \\
&+ (1-\delta_{n,0})(1-\delta_{n,1})(1-\delta_{n,2})\sum_{b=3}^{n}{\sum_{\{n_i\}}}'\frac{p_k(n_1|k-1)}{1-p_k(0|k-1)}\times\prod_{i=2}^{b-1}\frac{p_k(n_i|k-1-\sum_{j=1}^{i-1}n_j)}{1-p_k(0|k-1-\sum_{j=1}^{i-1}n_j)} \\
&\times \frac{1}{1-p_k(0|k-1-\sum_{j=1}^{b-1}n_j)}\left[q_k(n_b|k-1-\sum_{j=1}^{b-1}n_j)+p_k(n_b|k-1-\sum_{j=1}^{b-1}n_j)\frac{q_k(0|k-1-n)}{1-p_k(0|k-1-n)}\right] \qquad (13)
\end{aligned}
$$

where $n_b = n - \sum_{j=1}^{b-1} n_j$ and

$${\sum_{\{n_i\}}}' \equiv \sum_{n_1=1}^{n-b+1} \cdots \sum_{n_i=1}^{n-b+i-\sum_{j=1}^{i-1}n_j} \cdots \sum_{n_{b-1}=1}^{n-1-\sum_{j=1}^{b-2}n_j} .$$

Now that we have the expression for $P_k(n|k-1)$, we can compute numerically the average number of secondary spreaders $\bar{n}_\infty$ that an early spreader with $k-1$ ignorant neighbors generates during her spreader period by plugging Eq. 13 into Eq. 10. In Figure. 5.1 (F) we display the phase space of the rumor model on the λ–α plane for $k = 7$. We can clearly see that the rumor can not spread globally if the spreading rate λ is below a certain value, the so-called critical value λ_c, determined by the value of the annihilation rate α. Similarly, given a spreading rate λ, if the annihilation rate is above its critical value α_c, then the rumor can not have a global outreach. In order to compare the analytical results with our simulations, we set $\alpha = 0.5$ and compute λ_c, obtaining $\lambda_c \approx 0.0995$, which is very close to the value $\lambda_c = 0.1 \pm 0.005$ obtained from simulations.

5.3 Discussion

Zanette [72] reports phase transitions in the rumor model on small-world networks. While fixing the contagion parameters, the study shows that there is a phase transition between a regime in which the rumor is confined to a small neighborhood of the initial spreader and a regime where it spreads to a finite fraction of the population in the thermodynamic limit [72]. The transition is shown to be dependent on the network randomness or local clustering. Similarly, our study shows the existence of a phase transition between an absorbing state and an active state depending on both the contagion parameters and the network architecture. The absorbing state is given by confinement of the contagion to a small neighborhood due to slow spreading compared to the annihilation process. The active state on the other hand is the global outreach of the rumor in the thermodynamic limit. While we have focused on the random contact networks where each node has the same degree $k = 7$, we have performed simulations on Erdős-Rényi random graphs with average degree $\bar{k} = 7$ and obtained similar results qualitatively (figures not shown). Hence, we believe that our conclusions are not restricted to the simple network architecture considered here.

6 Scheduled milestones

The following dates and milestones are preliminary and will be adjusted as progress is made:

January 2014	Proposal presentation
January 2014 - February 2014	Paper writing on Section 4
January 2014 - March 2014	Work detailed on Section 3
February 2014 - March 2014	Dissertation writing
April 2014	Dissertation defence

References

[1] http://www.cdc.gov/flu/weekly/, 2013.

[2] http://ais.paho.org/phip/viz/ed_flu.asp, 2013.

[3] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting Flu Trends using Twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707, 2011.

[4] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[5] M. Ángeles Serrano, M. Boguñá, and A. Vespignani. Patterns of dominant flows in the world trade web. *J. Econ. Interac. Coord.*, 2:111–124, 2007.

[6] APS. Data sets for research, 2010.

[7] D. Balcan, V. Colizza, B. Gon¸calves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489, 2009.

[8] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of Computational Science*, 1(3):132–145, 2010.

[9] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Broeck, V. Colizza, and A. Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7(1):45, 2009.

[10] D. Balcan and A. Vespignani. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat Phys*, 7:581–586, 2011.

[11] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[12] M. Batty. The Geography of Scientific Citation. *Environ Plan A*, 35:761–765, 2003.

[13] C. Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68:314–316, 2007.

[14] M. Boguná, C. Castellano, and R. Pastor-Satorras. Langevin approach for the dynamics of the contact process on annealed scale-free networks. *Physical Review E*, 79:036110, 2009.

[15] K. Börner, S. Penumarthy, M. Meiss, and W. Ke. Mapping the Diffusion of Information Among Major U.S. Research Institutions. *Scientometrics*, 68:415–426, 2006.

[16] L. Bornmann, L. Leydesdorff, C. Walch-Solimena, and C. Ettl. Mapping excellence in the geography of science: An approach based on Scopus data. *Journal of Informetrics*, 5(4):537–546, 2011.

[17] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comp. Net. ISDN Sys.*, 30:107, 1998.

[18] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE*, 6(7):e20814, 2011.

[19] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71:027103, 2005.

[20] Center for International Earth Science Information Network (CIESIN) Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). The Gridded Population of the World Version 3 (GPWv3): Population Grids. `http://sedac.ciesin.columbia.edu/gpw`, 2005.

[21] D. Centola and M. Macy. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology*, 113(3):702–734, 2007.

[22] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1:8–15, 2007.

[23] G. Chowell, M. Miller, and C. Viboud. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiology & Infection*, 136:852–864, 6 2008.

[24] F. Ciulla, D. Mocanu, A. Baronchelli, B. GonÃğalves, N. Perra, and A. Vespignani. Beating the news using social media: the case study of American Idol. *EPJ Data Science*, 1(1):1–11, 2012.

[25] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2015–2020, 2006.

[26] V. Colizza and A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.*, 99:148701, 2007.

[27] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of 1st Workshop on Social Media Analytics (SOMA '10)*, 2010.

[28] D. J. Daley and D. G. Kendal. Stochastic rumours. *IMA Journal of Applied Mathematics*, 1(1):42–55, 1965.

[29] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Critical phenomena in complex networks. *Review of Modern Physics*, 80:1275, 2008.

[30] A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman. Influenza forecasting with google flu trends. *PLoS ONE*, 8(2):e56176, 2013.

[31] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69:131–152, 2006.

[32] J. D. Frame, F. Narin, and M. P. Carpenter. The Distribution of World Science. *Social Studies of Science*, 7:501–516, 1977.

[33] K. Frenken, S. Hardeman, and J. Hoekman. Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, 3:222–232, 2009.

[34] E. Garfield. Citation Analysis as a Tool in Journal Evaluation. *Science*, 178:471–479, 1972.

[35] GeoNames. Geonames. http://www.geonames.org/, Retr. 2012.

[36] G. Ghasemiesfeh, R. Ebrahimi, and J. Gao. Complex contagion and the weakness of long ties in social networks: Revisited. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pages 507–524, 2013.

[37] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, Feb. 2009.

[38] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[39] M. S. Granovetter. The strength of weak ties: A network theory revisited. *Sociological theory*, 1:201–233, 1983.

[40] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci.*, 102:16569–16572, 2005.

[41] J. E. Hirsch. Does the h index have predictive power? *Proc. Natl. Acad. Sci.*, 104:19193–19198, 2007.

[42] H. Horta and F. Veloso. Opening the box: comparing EU and US scientific output by scientific field . *Technological Forecasting & Social Change*, 74:1334–1356, 2007.

[43] M. Karsai, K. Kaski, and J. Kertész. Correlated Dynamics in Egocentric Communication Networks. *PLoS ONE*, 7:e40612, 2012.

[44] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83:025102, 2011.

[45] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animal*. Princeton University Press, 2008.

[46] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London*, 115:700–721, 1927.

[47] D. K. King. The scientific impact of nations. *Nature*, 430:311–316, 204.

[48] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[49] G. Krings, M. Karsai, S. Bernhardsson, V. Blondel, and J. Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):1–16, 2012.

[50] R. Lambiotte, V. D. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. V. Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317 – 5325, 2008.

[51] L. Leydesdorff and P. Zhou. Are the contributions of China and Korea upsetting the world system of science? *Scientometrics*, 63:617–630, 2005.

[52] J. Li and C. Cardie. Early Stage InïňĆuenza Detection from Twitter. http://arxiv.org/abs/1309.7340.

[53] D. P. Maki and M. Thompson. *Mathematical models and applications: with emphasis on the social, life, and management sciences*. Prentice-Hall, 1973.

[54] J. Marro and R. Dickman. *Nonequlibrium Phase Transition in Lattice Models*. Cambridge University Press, 1999.

[55] R. M. May. The Scientific Wealth of Nations. *Science*, 7:793–796, 1997.

[56] A. Mazloumian, D. Helbing, S. Lozano, R. P. Light, and K. Börner. Global multi-level analysis of the 'scientific food web'. *Scentific Reports*, 3:1167, 2013.

[57] Y. Moreno, M. Nekovee, and A. F. Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69:066130, 2004.

[58] F. Narin and M. P. Carpenter. National Publication and Citation Comparisons. *Journal of the American Society for Information Science*, 26:80–93, 1975.

[59] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374:457–470, 2007.

[60] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[61] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9(10):e1003256, 2013.

[62] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, J. Kertész, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9:179, 2007.

[63] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS*, 104:7332, 2007.

[64] R. K. Pan, K. Kaski, and S. Fortunato. World citation and collaboration networks: uncovering the role of geography in science. *Scentific Reports*, 2:902, 2012.

[65] S. Redner. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.

[66] I. Rodríguez-Iturbe and A. Rinaldo. *Fractal River Basins: Chance and Self-Organization*. Cambridge University Press, 2001.

[67] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[68] D. Stauffer and A. Aharony. *Introduction to Percolation Theory*. Taylor & Francis, 2003.

[69] M. Tizzoni, P. Bajardi, C. Poletto, J. Ramasco, D. Balcan, B. Goncalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC Medicine*, 10(1):165, 2012.

[70] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

[71] A. Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nat Phys*, 8:32–39, 2012.

[72] D. H. Zanette. Critical behavior of propagation on small-world networks. *Physical Review E*, 64:050901, 2001.

[73] J. Zhao, J. Wu, and K. Xu. Weak ties: Subtle role of information diffusion in online social networks. *Physical Review E*, 82:016105, 2010.