Disease surveillance on complex social networks

J. L. Herrera,[1,2] Ravi Srinivasan,[3] Alison Galvani,[4] and Lauren Ancel Meyers[1]

[1]*Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA*
[2]*Departamento de Cálculo, Escuela Básica de Ingeniería, Universidad de Los Andes, Mérida, Venezuela*
[3]*Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, USA*
[4]*School of Medicine, Yale University, CT, New Haven, 06510, USA*

(Dated: April 27, 2015)

As infectious disease surveillance systems expand to include digital, crowd-sourced, and social network data, public health agencies are gaining unprecedented access to high-resolution data and have an opportunity to selectively monitor informative individuals. Contact networks, which are the webs of interaction through which diseases spread, determine whether and when individuals become infected, and thus who might serve as early and accurate surveillance sensors. Here, we evaluate three strategies for selecting sensors—sampling the most connected, random, and friends of random individuals—in three complex social networks—a simple scale-free network, an empirical Venezuelan college student network, and an empirical Montreal wireless hotspot usage network. Across five different surveillance goals—early and accurate detection of epidemic emergence and peak, and general situational awareness—we find that the optimal choice of sensors depends on the public health goal, underlying network and reproduction number of the disease (R_0). For slowly spreading diseases (low R_0), the most connected individuals provide the earliest and most accurate information about both the onset and peak of an outbreak. However, identifying network hubs is often impractical, and they can be misleading if monitored for general situational awareness, if the underlying network has significant community structure, or if R_0 is high or unknown. The friends-of-random strategy offers a more practical and robust alternative. It can be readily implemented without prior knowledge of the network, and by identifying sensors with higher than average, but not the highest, epidemiological risk, it provides reasonably early and accurate information. Taking a theoretical approach, we also derive the optimal surveillance system for early outbreak detection but find that real-world identification of such sensors would be nearly impossible.

I. INTRODUCTION

Public health agencies rely on diverse sources of information for detecting emerging outbreaks, situational awareness (e.g., estimating prevalence or severity), prediction of future burden, and triggering effective control measures. For influenza alone, the CDC has deployed approximately eight different surveillance systems [1]. With public health facing increasing budget constraints [2], Meaningful Use soon providing public health access to volumes of real-time electronic hospital records, and the explosion of internet-source data [3], disease surveillance is at a critical juncture.

HealthMap—a website that aggregates worldwide news to generate global health risk map—was among the first effective internet-driven surveillance systems [4, 5]. In 2009, Google Flu Trends—a detection algorithm for internet search queries of influenza-related terms—brought next-generation surveillance to the forefront of public health [6–11]. It correlated with seasonal dynamics in the US and Europe, but fell short during the 2009 H1N1 pandemic [12–14]. Next-generation surveillance has since exploded with efforts to harness data from search engines [15, 16], crowdsourcing (e.g., Flu Near You in the US and Influenzanet in Europe) [17, 18], Twitter (e.g., MappyHealth) [19, 20], and Facebook [21, 22].

While promising, public health agencies face the significant challenge of effectively integrating these new data sources to achieve specific surveillance objectives. Many next generation data sources, whether passively scraping data intended for another purpose or actively engaging volunteer participants, are collecting data from *networked* systems. The data providers have physical or virtual connections to each other through which disease, opinions or information spreads.

Decades of sociology and epidemiology research have demonstrated that the resulting network structure can profoundly influence the spread of disease and behavior, and determine if and when individuals are affected [23–30]. In particular, there are diverse methods for quantifying the importance or *centrality* of a *node* (individual) in a network, many of which can have been shown to predict epidemiological risk and indicate optimal targets for interventions such as vaccination [31–33, 33–37, 39].

In designing disease surveillance systems for networked populations, one seeks to identify nodes (*sensors*) that are likely to provide early and accurate indications of epidemic activity. While superficially similar to the problem of selecting optimal targets for vaccination, the best sensors are not necessarily those most likely to be infected and infect others. Such nodes may be the earliest or most often infected, but be unreliable indicators of the broader epidemiological situation. Conversely, a completely representative cross-section of a network may provide real-time situational awareness, but be too slow for triggering effective control measures. For livestock diseases, Bajardi et al. developed a network path based strategy for identifying surveillance locations that should provide timely and accurate outbreak data [38]. Christakis et

al. performed an experimental comparison of two social-network-based strategies in a college population [42]. In one strategy, the sensors were a random selection of students; in the other, the sensors were identified as friends of one or more random students. The *friends-of-random* surveillance group was expected be biased towards more central individuals, and provided almost two-weeks earlier indication of the 2009-2010 pandemic H1N1 influenza epidemic relative to the *random* surveillance group.

Here, we use a mathematical model to systematically evaluate these and other strategies for selecting surveillance sensors, across several networks and an ensemble of common public health objectives. We quantify the timing and accuracy of the information gained by monitoring the disease states of strategically chosen sensors, as well as the robustness of the information across epidemiological scenarios (that is, different values of the disease reproduction number, R_0). As hypothesized, the best surveillance targets are not always those with the highest epidemiological risk or those most representative of the underlying network. While not optimal for all public health objectives, the friends-of-random strategy balances risk and representativeness, provides reasonably robust, accurate and early warning, and can be applied without knowledge of the underlying contact network.

II. METHODS

A. Epidemic model

We simulate disease outbreaks in contact networks using a simple stochastic chain-binomial susceptible-exposed-infected-recovered (SEIR) simulation model [40, 41]. Networks consist of *nodes* representing individuals and *edges* between pairs of nodes representing contacts between individuals. The *degree* of a node is the number of other nodes to which it is connected via an edge.

During a simulated epidemic, each node is in one of four states: susceptible (S), exposed to disease but not yet infectious (E), infectious (I), or recovered (R). If a node i in state S shares an edge with a node j in state I, then j will infect i with probability β and i will transition from S to E. After a period of l days, i will enter the infectious state I. It will remain infectious for d days, and then move to the immune state R.

The *reproduction number* of a disease, denoted R_0, indicates the growth rate of an epidemic and the expected number of secondary infections produced by a single infected host in an entirely susceptible population. Large epidemics are only possible when $R_0 > 1$. In a random network, R_0 is related to β as follows [44]:

$$R_0 = \beta \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right), \tag{1}$$

where $\langle k \rangle$ and $\langle k^2 \rangle$ are the mean degree and the mean squared degree, respectively, of nodes in the network. R_0 depends explicitly on both the intrinsic transmission rate of the pathogen and the structure of the network. For our analyses, we specify R_0 and use equation 1 to solve for the corresponding β. For the empirical networks considered, clustering, modularity and other non-random structures may cause the resulting R_0 to differ slightly from the one initially specified.

For each simulation, we fix the latent period to $l = 4$ days and the infectious period to $d = 7$, roughly in the range of estimates for common respiratory diseases, including influenza [45, 46]. Epidemics are initialized with a single random infected node and allowed to evolve until there are no remaining infected nodes.

B. Contact networks

Social interactions often give rise to complex network structures, with features that impose non-trivial constraints on the flow of information, behavior and disease [47–50]. We evaluated network-based surveillance strategies using three classes of social networks with distinct topological attributes.

1. *Scale-free networks*: Networks generated using the Barabasi-Albert algorithm [50] with $N = 10,000$ nodes, starting with $m_0 = 3$ nodes and iteratively adding nodes with edges to $m = 3$ existing nodes.

2. *Student network*: A social network formed by $N = 4,634$ students (nodes) of the Engineering Department from Universidad de Los Andes in Merida - Venezuela, where edges indicate that students took the same class during the fall 2008 semester.

3. *Montreal WiFi Network*: A co-location network for $N = 103,425$ users (nodes) of the Île Sans Fil free public wireless network in Montreal, Canada, where edges represent concurrent hotspot usage.

The degree distributions of the scale-free and Montreal network resemble power laws [50, 51], while the student network has a relatively homogeneous (Poisson) degree distribution. The Montreal network, but not the other two, exhibits strong community structure [51]

C. Surveillance strategies

We propose three strategies for designing network-based surveillance systems. Each strategy is a criteria for selecting a subset of individuals to monitor for their disease state: (1) *most connected*: select the highest degree individuals in the network; (2) *random*: select individuals at random; and (3) *random acquaintance*: select a random acquaintance of random individuals (which should be biased toward high degree individuals [52]). These

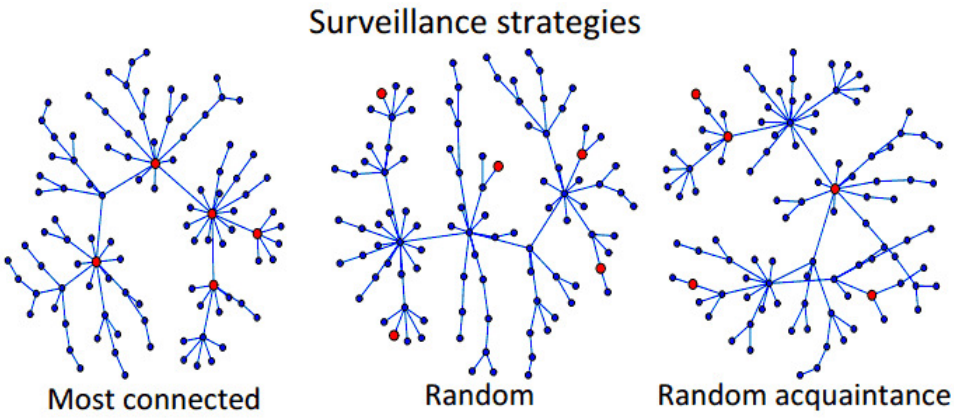

FIG. 1: Schematic representation of the proposed surveillance strategies. Red nodes are selected for surveillance.

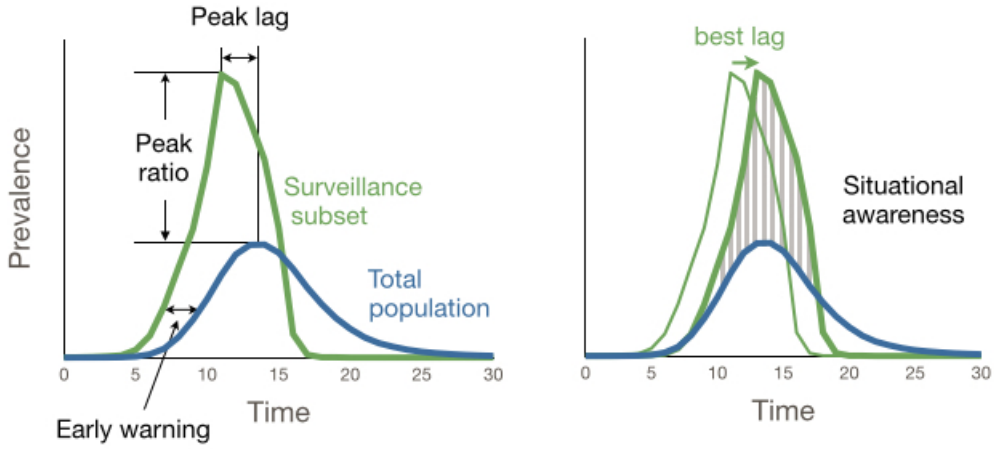

FIG. 2: Surveillance objectives. To evaluate strategies, we compare the epidemic curve (prevalence time series) of the subset of nodes under surveillance (green lines) with the epidemic curve for the whole population (blue lines). We calculate the time lag between the surveillance group and whole population reaching 7% prevalence, except for the Montreal network, where we use 1% prevalence (*early warning*), the time lag between the surveillance group and whole population reaching their epidemic peaks and the ratio of the magnitudes of the two peaks (*peak forecasting*), and the complement of the normalized mean absolute error (MAE) (*situational awareness*).

strategies are illustrated in Figure 1 for a scale-free network, where each surveillance subset includes five of the 100 nodes (in red).

The most connected strategy assumes complete knowledge of the network structure, whereas the random and random acquaintance strategies do not.

D. Evaluation of surveillance strategies

We assess the performance of each surveillance strategy with respect to several different public health goals (Figure 2). For each strategy-network combination, we build surveillance subsets by selecting 1% of all nodes (unless otherwise specified) via the strategy. We then estimate performance by running stochastic SEIR simulations, and make the following four comparisons between the prevalence time-series in the whole population to that of surveillance subset:

1. *Early warning*: The lag between the surveillance subset reaching 7% prevalence and the entire population reaching 7% prevalence (except for the Montreal network, where we use 1%).

2. *Peak timing*: The lag between the surveillance subset reaching its epidemic peak and the entire population reaching its epidemic peak.

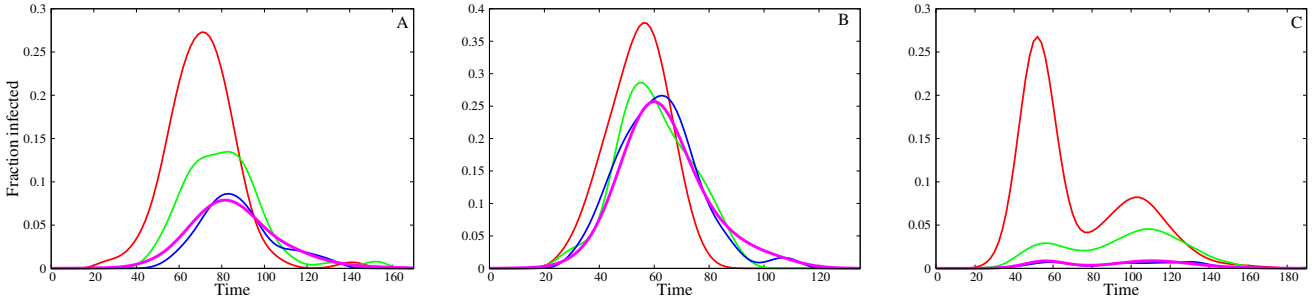

FIG. 3: Typical epidemic curves for the three focal networks: (A) scale-free, (B) student and (C) Montreal. Lines indicate the fraction of infected nodes overall (magenta) and in 1% subsets of nodes selected according to the most connected (red), random (blue), and random acquaintance (green) surveillance strategies during a single SEIR simulation with $R_0 = 3$.

3. *Peak magnitude*: The ratio of peak prevalence in the surveillance subset and peak prevalence overall.

4. *Situational awareness*: The complement of the normalized mean absolute error (MAE), minimized over possible lags, given by

$$1 - \min_{\lambda} \frac{\sum_t \left| \frac{x_t}{M} - \frac{y_{t+\lambda}}{N} \right|}{\sum_t \left(\frac{x_t}{M} + \frac{y_{t+\lambda}}{N} \right)}. \qquad (2)$$

Here, x_t and y_t are the prevalence in the surveillance subset and whole population at time t, respectively, N is the population size, M is the size of the surveillance subset, and λ is the lag.

All results are averaged over 200 stochastic SEIR simulations. At the beginning of each simulation, the surveillance subset is chosen anew according to the given strategy. For each objective function, we quantify both the magnitude of the effect and its stability with respect to a key epidemiological quantity, R_0. If the information provided by a surveillance system is highly sensitive to R_0, then it may be unreliable or uninterpretable in situations where R_0 is unknown or changing.

III. RESULTS

In all three networks, the most connected strategy selects subsets of nodes that tend to experience earlier and more intense epidemics, and the random strategy yields collections of sensors that almost perfectly resemble the population as a whole (Figure 3). The random acquaintance strategy produces subsets that provide some early warning in the scale-free and Montreal networks, but not in the highly homogeneous student network. The epidemic curves in the Montreal network occasionally exhibit multiple peaks, driven by underlying community structure [51].

A systematic evaluation of the three strategies in the three focal networks (Figure 4) confirms some of the patterns observed anecdotally (Figure 3). The most connected strategy consistently provides the earliest warning for both the beginning and peak of the season, and exhibits the highest peaks and the least overall similarity to the full epidemic curve (Figure 4, red points). However, the duration of the early warning can be highly sensitive to R_0, presenting a challenge when there is uncertainty regarding R_0. For example, when $R_0 = 3$, the most connected surveillance subset crosses the season onset threshold an average of 6.93, 25.89 and 33.51 days before the entire population in the student, scale-free and Montreal networks respectively; when $R_0 = 5$, these early warning periods decrease to averages of 3.92, 13.7 and 17.21 days, respectively (Figure 4B, 4D and 4F). The epidemic peak in the most connected surveillance subsets also depends on R_0, impeding estimation of peak burden under uncertainty (Figure 4A, 4C and 4E). In the Montreal network, the average ratio between the peak in the surveillance subset and the peak overall decreases from 27.57 to 14.67 as R_0 increases from 3 to 5 (Figure 4E). In general, as R_0 increases, the height of the epidemic peak in the entire population approaches that in the most connected subset of the population.

The random strategy yields surveillance systems that closely reflect the overall epidemiological dynamics, with early warning values close to zero and peak ratios close to one, across all networks and values of R_0 (Figure 4, blue points). The random acquaintance surveillance groups perform relatively well in both the scale-free and Montreal networks (Figures 4A, 4B, 4E and 4F, green points). It offers some degree of early warning for both season onset and peak, though not as much as the most connected group, while, importantly, exhibiting greater robustness with respect to R_0 in the timing of early warning, peak ratio, and situational awareness (overall correlation between surveillance epidemic curve and population epidemic curve). However, in the student network, the random and random acquaintance subsets are virtually indistinguishable (Figures 4C and 4D). This network is highly homogeneous, with most nodes having close to

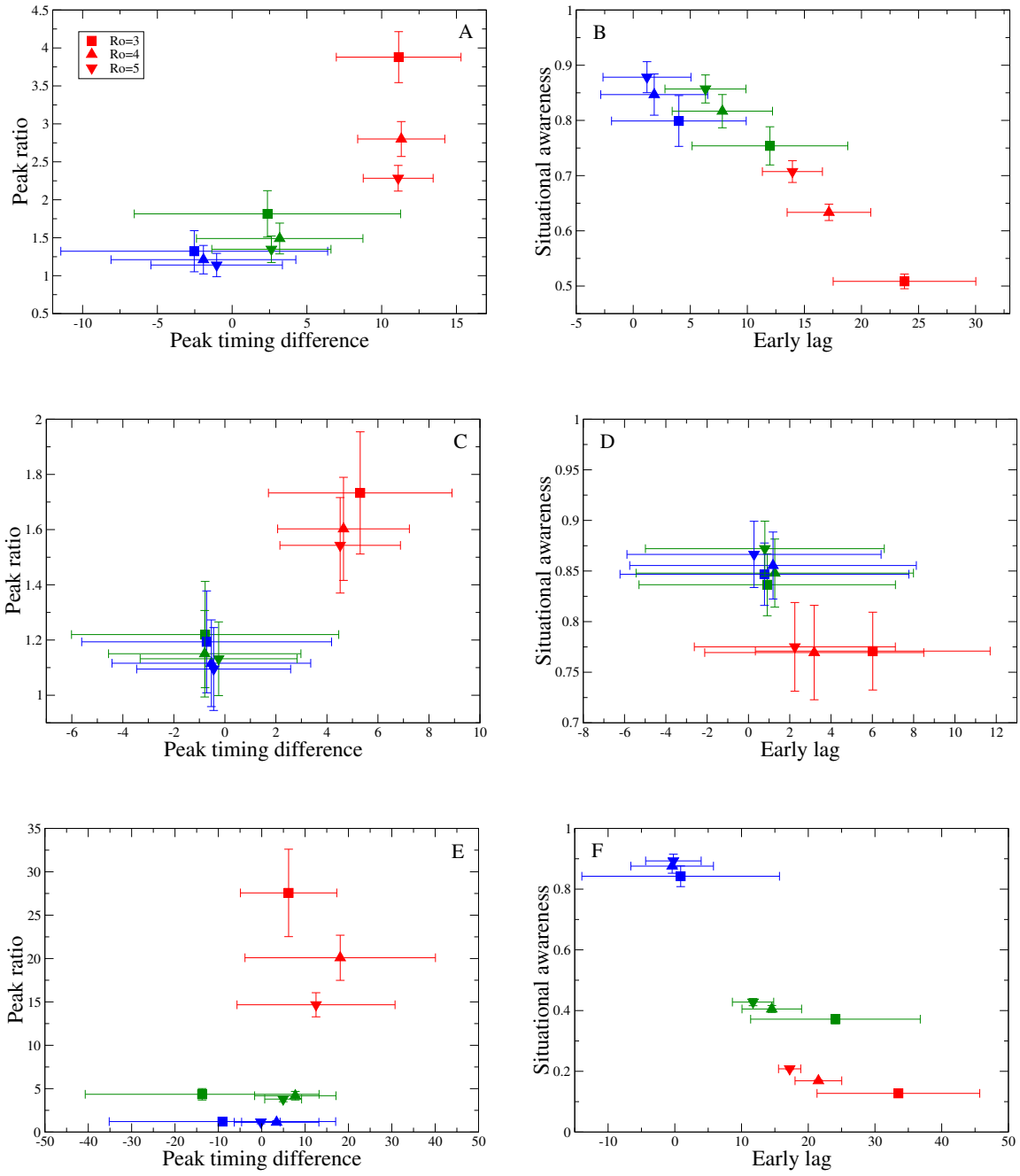

FIG. 4: Performance of most connected (red), random (blue), and random acquaintance (green) strategies with respect to predicting the timing and magnitude of the peak (graphs A, C, and E), and achieving early warning and situational awareness (graphs B, D, and F). Points and error bars indicate mean and standard deviation in performance over 200 simulations, respectively. Performance depends on both R_0 and network structure: scale-free (graphs A and B), student (graphs C and D), and Montreal (graphs E and F).

the average number of contacts. Thus, random acquaintances tend to be average as well.

As the size of a surveillance system increases, the detected epidemic curves become more similar to the full epidemic curve, thereby providing better situational awareness (Figure 5). In the scale-free and student networks, the performance of the three different surveillance strategies stabilizes to a quasi-stationary state around 3%, which entails tracking 300 of 10,000 nodes and 139 of 4,634 nodes in the two networks, respectively. In the

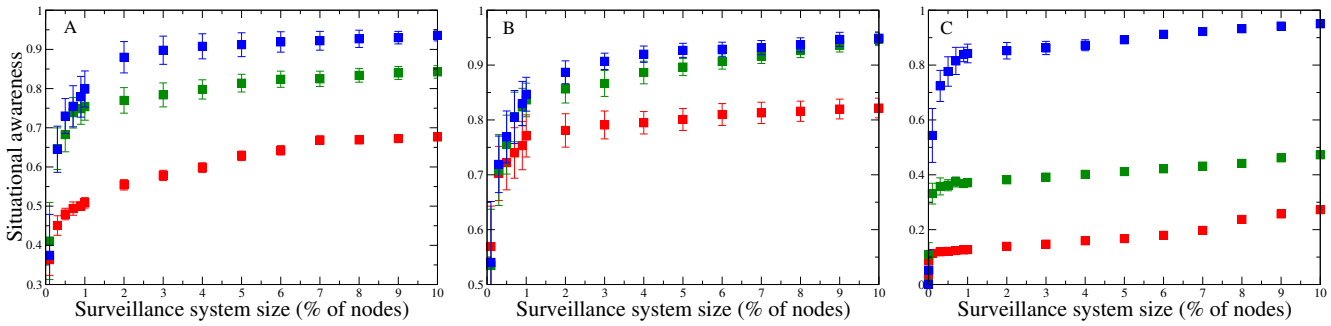

FIG. 5: Size of surveillance systems impacts performance. Situational awareness (similarity between surveillance epidemic curve and full epidemic curve) improves as the surveillance system grows in the (A) scale-free, (B) student and (C) Montreal networks. Surveillance groups were chosen using the most connected (red), random (blue), and random acquaintance (green) strategies.

Montreal network, the random and random acquaintance groups reach their optimal performance by 0.5% (517 of 103,425 of nodes), while the most connected group is still improving even at 10% (10,342 of 103,425 of nodes).

There are innumerable alternative strategies for selecting surveillance nodes, including prioritization based on other well-studied network centrality measures. For example, k-shell decomposition [37] and eigenvector centrality [48] are more computationally demanding and challenging to implement in practice, yet are not expected to significantly improve outcomes (Figure S2).

A theoretically optimal surveillance strategy

Following Newman [48], we use percolation theory to model SIR epidemics on networks, and derive the optimal surveillance group for early detection of an epidemic. Consider a disease with transmissibility β and recovery rate γ spreading through a network of size N. During the initial outbreak, the probabilities of each node being infected at time t are approximately given by the vector

$$\mathbf{x}(t) = e^{(\beta\kappa-\gamma)t}\mathbf{v}, \tag{3}$$

where κ is the leading eigenvalue of the adjacency matrix and $\mathbf{v}$ its corresponding eigenvector [48].

We extend this equation to model the time lag between a subset S of the network of size $M \leq N$ reaching a given prevalence threshold p and the overall population prevalence reaching p. Let $\mathbf{1}$ be the vector of length N containing all ones, $\mathbf{1} = (1, \ldots, 1)$, and $\mathbf{1}_S$ be the binary vector of dimension N indicating which M nodes are under surveillance

$$\mathbf{1}_S = \begin{cases} 1 & \text{if node } i \text{ is in the surveillance subset } S \\ 0 & \text{otherwise.} \end{cases}$$

For example, if the 1% most connected nodes were selected for surveillance in a network of size $N = 1000$, then the entries of $\mathbf{1}_S$ corresponding to the ten highest degree nodes would be one, and the remaining entries would be zero.

Let τ and τ_S be the times at which the entire population and a given surveillance group reach the prevalence threshold p, respectively. Substituting into the above equation, we find

$$p = \frac{\mathbf{x}(\tau)\cdot\mathbf{1}}{N} = \frac{e^{(\beta\kappa-\gamma)\tau}\mathbf{v}\cdot\mathbf{1}}{N} \tag{4}$$

and

$$p = \frac{\mathbf{x}(\tau_S)\cdot\mathbf{1}_S}{M} = \frac{e^{(\beta\kappa-\gamma)\tau_S}\mathbf{v}\cdot\mathbf{1}_S}{M}. \tag{5}$$

To solve for the length of early warning achieved through surveillance $\Delta\tau = \tau_S - \tau$, we combine these as

$$\frac{e^{(\beta\kappa-\gamma)\tau}\mathbf{v}\cdot\mathbf{1}}{N} = \frac{e^{(\beta\kappa-\gamma)\tau_S}\mathbf{v}\cdot\mathbf{1}_S}{M}. \tag{6}$$

This implies

$$\Delta\tau = \frac{1}{\beta\kappa-\gamma}\ln\left(\frac{c}{c_S}\right), \tag{7}$$

where $c = \mathbf{v}\cdot\mathbf{1}/N$ and $c_S = \mathbf{v}\cdot\mathbf{1}_S/M$ are the average eigenvector centralities in the network as a whole and the surveillance subset, respectively. The early season lag between the surveillance subset and the whole population can thus be positive or negative, and depends on ratio of their average eigenvector centralities.

We assessed the validity of this mean field approximation by comparing the expected early warning period (Equation 7) to simulated early warning periods for both the most connected subset and the subset of the 1% highest eigenvector centrality nodes. To match the assumptions of our mean field model, we simulated SIR rather than SEIR transmission dynamics. The simulations mirrored the theoretical expectations for both types

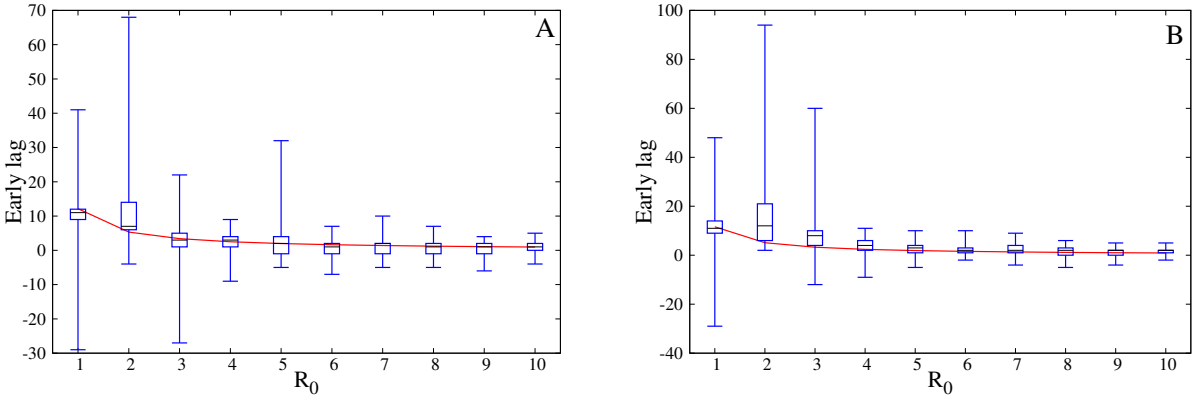

FIG. 6: Comparing theory and simulation for early warning period in the scale free network. As R_0 increases, the lag between the surveillance subset and the entire population reaching the early detection threshold decreases for both the (A) 1% highest eigenvector centrality nodes and (B) 1% highest degree nodes. Red curves indicate theoretical approximations; box plots show distribution of SIR simulation results.

of surveillance subsets in all three networks, as shown for the scale free network (Figure 6).

Next, we solve for the surveillance subset that maximizes the length of the early warning period. For a given surveillance system size M, the earliest warning is achieved when $\mathbf{1}_S$ indicates the M nodes in the network with the highest valued entries in $\mathbf{v}$. In other words, the theoretically optimal surveillance strategy for early warning of epidemic onset is selecting nodes with the highest eigenvector centrality.

Importantly, $\Delta\tau$ depends on the disease parameters β and γ. Regardless of the choice of surveillance nodes $\mathbf{1}_S$, the length of the early warning period will, therefore, increase as R_0 decreases. The only exception is when the average eigenvector centrality in the surveillance subset equals that in the population as a whole ($c = c_S$). In that case, there is no early warning ($\Delta\tau = 0$). These conclusions are reflected in the sensitivity to R_0 observed in our simulations (Figures 4B, 4D and 4F).

For the networks under consideration, the most connected strategy produces surveillance groups with relatively high eigenvector centrality while the random strategy yields groups with average eigenvector centrality. However, eigenvector centrality in random acquaintance groups depends on the underlying network: in homogeneous networks like the student network, it will be average; in heterogeneous networks, it will be above average.

Finding the optimal surveillance nodes

Identifying individuals with the highest eigenvector centrality is quite difficult in real-world populations, where the large-scale network structure is generally unknown. On the other hand, finding individuals with above average degree centrality is possible using local information. If eigenvector centrality is correlated to degree centrality, as it is in the three networks we consider (see Figure 7), it may be possible to use highly connected nodes as a proxy for high eigenvector centrality nodes.

One strategy for finding high degree centrality nodes is to follow chains of random acquaintances. This has been explored extensively in the context of respondent-driven sampling, such as chain-referral (i.e., "snowball") sampling [59]. In particular, consider the simple random walk in which, at each step, the walker moves to a neighboring node selected uniformly at random. For connected, undirected networks, this is equivalent to the PageRank algorithm with no damping factor [48]. Assuming the network is fully connected, the distribution of the random walker after m steps approaches a stationary distribution as $m \to \infty$, in which the probability of landing on a node is exactly proportional its degree [60]. Thus, the more connected the node, the more likely we are to reach it.

Precisely, let k_i be the degree of node i and $P(k)$ the degree distribution of the network. The nth moment of the degree distribution is

$$\langle k^n \rangle = \frac{1}{N} \sum_i k_i^n = \sum_k k^n P(k). \tag{8}$$

Let D_m denote the degree of the node at which the random walk resides on the mth step, starting from a node chosen uniformly at random. Assuming the mean degree $\langle k \rangle < \infty$, then the distribution of D_∞ is given by $kP(k)/\langle k \rangle$. If $\langle k^3 \rangle < \infty$, which is true for any finite graph but will be violated for power-law networks without cutoff, the mean and standard deviation of D_∞ are

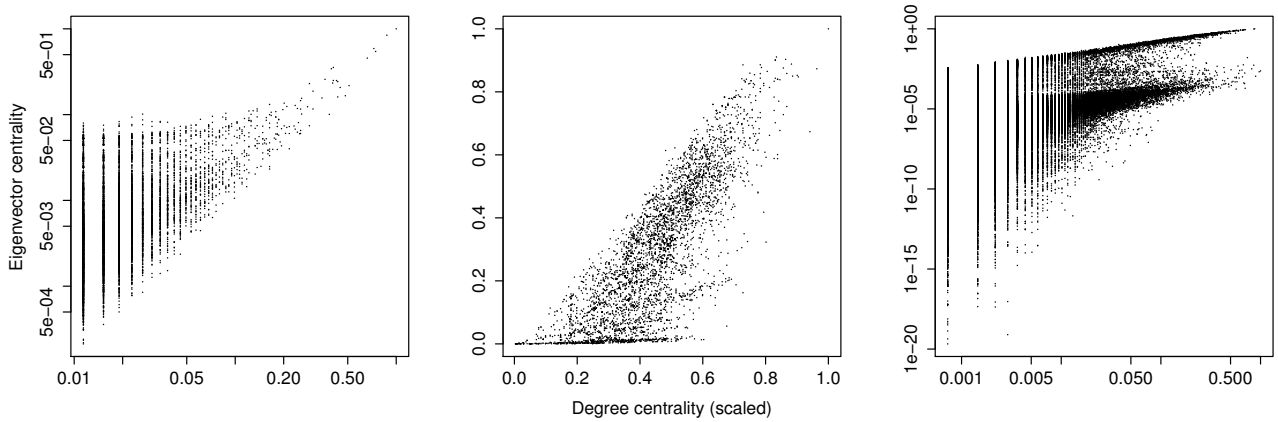

FIG. 7: Scatter plots of eigenvector centrality vs. (scaled) degree centrality of nodes in the (A) scale-free, (B) student, and (C) Montreal networks. Both eigenvector and degree centralities are scaled to have maximum value 1, and log-log plots are shown for (A) and (C). The student network shows strong correlation between the two centrality measures, with a Spearman rank correlation coefficient of 0.819, while for the scale-free and Montreal networks the measures have more moderate rank correlation coefficients of 0.441 and 0.620, respectively.

given by

$$\mu_\infty = \frac{\langle k^2 \rangle}{\langle k \rangle}, \qquad \sigma_\infty = \sqrt{\frac{\langle k^3 \rangle}{\langle k \rangle} - \left(\frac{\langle k^2 \rangle}{\langle k \rangle} \right)^2}. \qquad (9)$$

By comparison, the distribution of randomly sampled nodes (D_0) has mean $\mu_0 = \langle k \rangle$ and standard distribution $\sigma_0 = \sqrt{\langle k^2 \rangle - \langle k \rangle^2}$. Thus, the random walk sample is biased towards nodes with larger degrees. For intermediate values of m, the distribution of D_m can only be derived with full knowledge of the underlying graph. Instead, note that this distribution converges to that of D_∞ at a rate that depends on the second largest eigenvalue of the adjacency matrix of the graph. If this eigenvalue is close to one, which is usually the case for connected networks with high modularity, convergence is very slow and random walk sampling may require many steps to achieve its optimal performance. Methods that bias the random walk towards higher eigenvector centrality nodes should be more effective in this setting. For example, the maximal entropy random walk (MERW) samples nodes proportional to eigenvector centrality just as the simple random walk considered earlier samples nodes according to their degree centralities [61]. However, MERW has transition probabilities that require global information about the network and is therefore impractical to implement without approximation as part of any sampling strategy.

Equations 9 provide a theoretical upper bound to the mean centrality that can be achieved when using a random walk on a network to design a surveillance system. In particular, for a random-walk surveillance subset of size $M = \epsilon N$ with fixed ϵ and N large, the empirical mean of the sample will become approximately normal with mean μ_∞ and standard deviation $\sigma_\infty/\sqrt{M}$, as illustrated for our three study networks (Figure 8).

IV. CONCLUSIONS

The success of both traditional surveillance systems like the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and next generation participatory systems like FluNearYou [17, 18], depends on targeted recruitment of reliable, informative providers. With Meaningful Use and the advent of digital disease detection, we are moving from an era of sparse, volunteer-based data into an era of data inundation [16, 56]. Yet, we still face the challenge of finding reliable data sources. Effective mining of electronic medical records, social media and other internet source data, such as Google, Twitter or Facebook, requires sifting through petabytes of data for streams that can provide early and accurate information about emerging outbreaks. While random representative sampling is a good rule-of-thumb and has guided the development of numerous surveillance systems, we can improve performance by exploiting our evolving understanding of social networks and their impacts on infectious disease dynamics [24, 28–30, 41, 44, 47–50, 57, 58].

In an ideal scenario where both the contact network and the reproduction number (R_0) of the disease are known in advance, public health agencies can monitor the most informative nodes and achieve very accurate

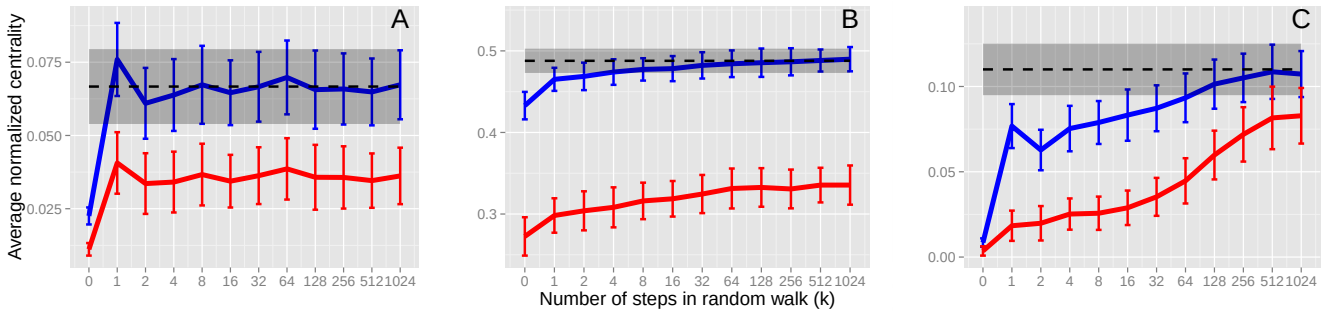

FIG. 8: Random walks increase centrality in the surveillance subset. For purposes of comparison, degree (blue) and eigenvector centrality (red) are divided by the maximum degree and maximum eigenvector centrality, respectively, in each network. Mean degree approaches its theoretical limit (dashed lines), and mean eigenvector centrality also increases as the random walks progress in the (A) scale-free (subset contains $\epsilon = 1\%$ of nodes), (B) student ($\epsilon = 2\%$), and (C) Montreal ($\epsilon = 0.1\%$) networks. As expected, the mean degree converges to a normal distribution with mean μ_∞ and standard deviation σ_∞ (gray shading) as k increases, where k is the number of steps in the walk. The random walks converge within a few steps in the scale-free and student networks, but require more steps in the highly modular Montreal network.

and early assessments of emerging epidemics. For example, we find that surveillance of the most connected individuals in the Montreal WiFi network can increase lead time on detecting epidemic emergence by two to three weeks and anticipating the epidemic peak by over a week. We show analytically that the optimal strategy for early detection of emerging outbreaks is targeting individuals with the highest eigenvector centrality, a measure that considers the connectivity of a node's neighbors, and those neighbors' neighbors, and so on [48]. It can only be calculated with full knowledge of the network, and estimates the proportion time spent on a node during an infinitely long random walk along the edges of the network. While providing the longest lead time (between the surveillance system crossing a prevalence threshold and the rest of the population crossing that threshold), the timing is highly dependent on R_0. In fact, regardless of which nodes are under surveillance, epidemiological activity becomes more synchronized and the lag time shrinks as R_0 increases.

However, this ideal scenario is unrealistic. When the contact network is unknown, we cannot easily identify the most central individuals, for almost any measure of centrality. Even if we could monitor the most connected individuals, correct interpretation of the resulting signal requires some knowledge of R_0. In general, low R_0 implies a longer lag time between epidemiological events in the surveillance group and corresponding events in the general population, and a larger discrepancy between prevalence in the surveillance group and overall epidemiological activity.

The random acquaintance strategy, which chooses random contacts of random nodes, provides a tractable method for identifying individuals with higher than average centrality. The intuition is that when choosing a random *friend of a node* rather than just a *random node*, the choice is biased towards individuals with more friends. In heterogeneous networks, like the scale-free and Montreal WiFi network considered here, random acquaintance groups provide some degree of early warning (significantly more than randomly selected nodes) and exhibit epidemic curves that reflect overall disease activity (significantly better than the most connected nodes). This is corroborated by the empirical finding that friends of random students served as better outbreak sentinels than random students during 2009 H1N1 pandemic [42]. Although the timing of the early warning and the discrepancy between the estimated prevalence and true prevalence will depend on R_0, the uncertainty can potentially be quantified and incorporated into confidence intervals.

In a relatively homogeneous network, like our Venezuelan student network, the random acquaintance strategy finds fairly average nodes and performs no better than the random strategy with respect to the surveillance objectives. This is consistent with basic theory on Erdős-Réyni networks: in a random network with a Poisson degree distribution, the average degree of random acquaintances will be exactly average [55]. Therefore, if a population is sufficiently homogeneous, surveillance systems should simply target random individuals or employ other methods for identifying highly connected individuals.

Acknowledgments

This research was supported by grant U01 GM087719 from NIGMS MIDAS.

[1] http://www.cdc.gov/flu/weekly/overview.htm.

[2] http://www.naccho.org/topics/infrastructure/lhdbudget/upload/Survey-Findings-Brief-8-13-13-3.pdf

[3] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917042/

[4] Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance Sans Frontires: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med 2008;5:e151-e151.

[5] Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008; 15:150-157

[6] Ginsberg J, Mohebbi MH, Patel RS, Branner L, Smolinski MS, et al. (2009), Detecting influenza epidemics using search engine query data. Nature 457:1012-1014.

[7] Carneiro HA, Mylonakis E (2009) Google trends: a web-based tool for real time surveillance of disease outbreaks. Clinical Infect. Dis. 49: 1557-1564.

[8] Fahad Pervaiz, Mansoor Pervaiz, Nabeel Abdur Rehman, Umar Saif; FluBreaks: Early Epidemic Detection from Google Flu Trends. J Med Internet Res. 2012 Sep-Oct; 14(5): e125.

[9] Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, et al. (2011) Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. PLoS ONE 6(4): e18687. doi:10.1371/journal.pone.0018687.

[10] Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, et al. (2013) Influenza Forecasting with Google Flu Trends. PLoS ONE 8(2): e56176. doi:10.1371/journal.pone.0056176.

[11] Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of google trends for epidemiological research: Lyme disease as an example. Geospatial Health. 2010;4:135-137.

[12] Valdivia A, Lpez-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks results for 2009-10. Euro Surveill. 2010;15(29):pii=19621.

[13] Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza surveillance networks. PLoS computational biology. 2012;8(4):e1002472.

[14] Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. Interpreting Google Flu Trends data for pandemic H1N1 influenza: The New Zealand experience. Euro Surveill. 2009;14(44):pii=19386.

[15] Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PloS one. 2013;8(5):e64323.

[16] Brownstein JS, Freifeld CC and Madoff LC, Digital Disease Detection Harnessing the Web for Public Health Surveillance, New England Journal of Medicine 21(360), 2153-2157 (2009).

[17] https://flunearyou.org/

[18] Chunara R, Aman S, Smolinski M, Brownstein J. Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. ISDS Conference Abstracts. 2013;5(1)

[19] Chew C, Eysenbach G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. PLoS ONE. 2010;5:e14118.

[20] Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PloS one. 2013;8(12):e83672.

[21] Boulos M, Sanfilippo A, Corley C, Wheeler S. Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine. 2010;100:16-23.

[22] Lee BK. Epidemiologic Research and Web 2.0the User-driven Web. Epidemiology. 2010;21(6):760-763.

[23] Newman, M.E.J., (2002), Spread of epidemic disease on networks. Phys. Rev. E 66, art. no.-016128.

[24] Meyers, L.A., Newman, M.E.J., et al., (2003), Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. Emerg. Infect. Dis. 9, 204.

[25] Salathé M, Jones JH (2010) Dynamics and Control of Diseases in Networks with Community Structure. PLoS Comput Biol 6(4): e1000736. doi:10.1371/journal.pcbi.1000736.

[26] Zonghua Liu and Bambi Hu; Epidemic spreading in community networks; Europhys. Lett., 72 (2), pp. 315321 (2005).

[27] Robert M. May and Alun L. Lloyd; Infection dynamics on scale-free networks; Phys Rev E, 64, 066112.

[28] Meyers, L.A., Pourbohloul, B., Newman M.E.J., Skowronskic D. M., Brunham, R.C.; Network theory and SARS: predicting outbreak diversity; Journal of Theoretical Biology 232 (2005) 7181.

[29] Meyers, L.A., Newman M.E.J., Pourbohloul, B.; Predicting epidemics on directed contact networks; Journal of Theoretical Biology 240 (2006) 400418.

[30] Volz, E.; SIR dynamics in random networks with heterogeneous connectivity; J. Math. Biol. (2008) 56:293310.

[31] Barrat, A., Barthélemy, M., Vespigniani, A. (2008) Dynamical processes on complex networks. Cambridge, UK: Cambridge University Press.

[32] Friedkin, N.E. (1991) Theoretical foundations for centrality measures. Am. J. Sociol. 96, 1478-1504.

[33] R. Albert, H. Jeong, A.-L. Barabási (2000), Error and attack tolerance of complex networks, Nature 406, 378482.

[34] Cohen, R., Erez, K., ben-Avraham, D., Havlin, S. (2001) Breakdown of the internet under intentional attack. Phys. Rev. Lett. 86, 3682-3685.

[35] Pastor-Satorras, R., Vespigniani, A. (2001) Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200-3203.

[36] Lloyd, A., May, R. (2001) Epidemiology: how viruses spread among computers and people. Science 292, 1316-1317.

[37] M. Kitsak, L. K. Gallos, S. Havlin, F. Lijeros, L. Muchnik, H. E. Stanley, H. A. Makse (2010) Identification of influential spreaders in complex networks, Nature 6, 888-893.

[38] Bajardi, P., Barrat, A., Savani, L., Colizza, V. (2012) Optimizing surveillance for livestock disease spreading through animal movements. J. R. Soc. Interface, 9, 2814-

2825.
[39] R. Cohen, S. Havlin, D. ben-Avraham (2003) Efficien immunization strategies for computer network and populations, Physical Review Letters 91, 247901.
[40] Abbey H; An examination of the Reed-Frost theory of epidemics. Hum Biol 1952, 24:201233.
[41] Ferrari MJ, Bansal S, Meyers LA, Bjornstad ON; Network frailty and the geometry of herd immunity; Proc R Soc B Biol Sci 2006, 273: 27432748.
[42] N.A. Christakis, J.H. Fowler (2010), Social Network Sensors for Early Detection of Contagious Outbreaks. PLoS ONE 5(9): e12948. doi:10.1371/journal.pone.0012948
[43] http://www0.health.nsw.gov.au/factsheets/guideline/pertusis.html
[44] Meyers, L.A. (2007) Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bulletin of the American Mathematical Society 44: 63-86.
[45] Huai Y, Xiang N, Zhou L, Feng L, Peng Z, Chapman RS, et al. Incubation period for human cases of avian influenza A (H5N1) infection, China, Emerging Infectious Diseases www.cdc.gov/eid Vol. 14, No. 11, November 2008.
[46] De Serres G, Rouleau I, Hamelin M-E, Quach C, Skowronski D, Flamand L, et al. Contagious period for pandemic (H1N1) 2009, Emerging Infectious Diseases www.cdc.gov/eid Vol. 16, No. 5, May 2010
[47] D.J. Watts and S.H. Strogatz; Collective dynamics of 'small-world' networks. Nature 393, 440-442 (1998).
[48] M.E.J. Newman (2010), Networks: An Introduction, Oxford University Press.
[49] A. Cho (2013), Network science at center of surveillance dispute, Science 340, 1272.
[50] A.-L. Barabási, R. Albert (1999), Emergence of scaling in random networks, Science 286, 509512.
[51] A.G. Hoen, T.J. Hladish, R.M. Eggo, M. Lenczner, A.P. Galvani, J.S. Brownstein and L.A. Meyers, Epidemic wave dynamics attributable to urban community structure (submitted to PLoS Computational Biology, April 2014)
[52] M.E.J. Newman (2001), Ego-centered networks and the ripple effect. Social Networks 25, 8395.
[53] J. D. Noh and H. Rieger (2004), Random Walks on Complex Networks, Phys. Rev. Lett. 92, 118701.
[54] Klemm K, Serrano MA, Eguiluz VM, San Miguel M (2012) A measure of individual role in collective dynamics: spreading at criticality. Scientific Reports 2:292.
[55] M. E. J. Newman, Ego-centered networks and the ripple effect, Social Networks 25, 8395 (2003).
[56] http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17207.pdf
[57] Bansal, S. and L.A. Meyers (2012) The impact of past epidemics on future disease dynamics. Journal of Theoretical Biology 309: 176-184.
[58] Volz, E.M., J.C. Miller, A.P. Galvani, L.A. Meyers (2011) Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics. PLoS Computational Biology 7: e1002042.
[59] Salganik, M. J., and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. Sociological methodology, 34(1), 193-240.
[60] Lovsz, L. (1993). Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty, 2(1), 1-46.
[61] Burda, Z., Duda, J., Luck, J. M., and Waclaw, B. (2009). Localization of the maximal entropy random walk. Physical Review Letters, 102(16), 160602.