Optimizing Provider Recruitment for Influenza Surveillance Networks

Samuel V. Scarpino[1]*, Nedialko B. Dimitrov[2], Lauren Ancel Meyers[1,3]

[1] The University of Texas at Austin, Section of Integrative Biology, Austin, Texas, United States of America, [2] Naval Postgraduate School, Operations Research Department, Monterey, California, United States of America, [3] The Santa Fe Institute, Santa Fe, New Mexico, United States of America

Abstract

The increasingly complex and rapid transmission dynamics of many infectious diseases necessitates the use of new, more advanced methods for surveillance, early detection, and decision-making. Here, we demonstrate that a new method for optimizing surveillance networks can improve the quality of epidemiological information produced by typical provider-based networks. Using past surveillance and Internet search data, it determines the precise locations where providers should be enrolled. When applied to redesigning the provider-based, influenza-like-illness surveillance network (ILINet) for the state of Texas, the method identifies networks that are expected to significantly outperform the existing network with far fewer providers. This optimized network avoids informational redundancies and is thereby more effective than networks designed by conventional methods and a recently published algorithm based on maximizing population coverage. We show further that Google Flu Trends data, when incorporated into a network as a virtual provider, can enhance but not replace traditional surveillance methods.

Citation: Scarpino SV, Dimitrov NB, Meyers LA (2012) Optimizing Provider Recruitment for Influenza Surveillance Networks. PLoS Comput Biol 8(4): e1002472. doi:10.1371/journal.pcbi.1002472

Editor: Mark M. Tanaka, University of New South Wales, Australia

Received August 9, 2011; **Accepted** February 29, 2012; **Published** April 12, 2012

Funding: This study was supported by the National Science Foundation DEB-0749097 to LAM and graduate research fellowship to SVS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: scarpino@utexas.edu

Introduction

Since the Spanish Flu Pandemic of 1918 – 1919, the global public health community has made great strides towards the effective surveillance of infectious diseases. However, modern travel patterns, heterogeneity in human population densities, proximity to wildlife populations, and variable immunity interact to drive increasingly complex patterns of disease transmission and emergence. As a result, there is an increasing need for effective, evidence-based surveillance, early detection, and decision-making methods [1–3]. This need was clearly articulated in 2009 by a directive from the Department of Homeland Security and the Centers for Disease Control and Prevention to develop a nationwide, real-time public health surveillance network [4,5].

The U.S. Outpatient Influenza-Like Illness Surveillance Network (ILINet) gathers data from thousands of healthcare providers across all fifty states. Throughout influenza season (CDC mandating reporting during weeks 40 – 20, which is approximately October through mid-May), participating providers are asked to report weekly the number of cases of influenza-like illness treated and total number of patients seen, by age group. Cases qualify as ILI if they manifest fever in excess of 100°F along with a cough and/or a sore throat, without another known cause. Although the CDC receives reports of approximately 16 million patient visits per year, many of the reports may use a loose application of the ILI case definition and/or may simply be inaccurate. The data are used in conjunction with other sources of laboratory, hospitalization and mortality data to monitor regional and national influenza activity and associated mortality. Similar national surveillance networks are in place in 11 EU countries and elsewhere around the globe [6–9].

Each US state is responsible for recruiting and managing ILINet providers. The CDC advises states to recruit one regularly reporting sentinel provider per 250,000 residents, with a state-wide minimum of 10 sentinel providers. Since 2003, the Texas Department of State Health Services (DSHS) has enrolled a total of 300 volunteer providers. Participating providers regularly drop out of the network; Texas DSHS aims to maintain approximately 200 active participants through year-round recruitment of providers in heavily populated areas (cities with populations of at least 100,000). DSHS also permits other (non-targeted) providers of family medicine, internal medicine, pediatrics, university student health services, emergency medicine, infectious disease, OB/GYN and urgent care to participate in the network. During the 2009 – 2010 influenza season, the Texas ILINet included 205 providers with approximately 50% reporting most weeks of the influenza season.

A number of statistical studies have demonstrated that ILI surveillance data is adequate for characterizing past influenza epidemics, monitoring populations for abnormal influenza activity, and forecasting the onsets and peaks of local influenza epidemics [10–16]. However, the surveillance networks are often limited by non-representative samples [17], inaccurate and variable reporting [12–14], and low reporting rates [6]. Some of these studies have yielded specific recommendations for improving the performance of the surveillance network, for example, inclusion of particular categories of hospitals in China [12], preference for general practitioners over pediatricians in Paris, France [14], and a

Author Summary

Public health agencies use surveillance systems to detect and monitor chronic and infectious diseases. These systems often rely on data sources that are chosen based on loose guidelines or out of convenience. In this paper, we introduce a new, data-driven method for designing and improving surveillance systems. Our approach is a geographic optimization of data sources designed to achieve specific surveillance goals. We tested our method by re-designing Texas' provider-based influenza surveillance system (ILINet). The resulting networks better predicted influenza associated hospitalizations and contained fewer providers than the existing ILINet. Furthermore, our study demonstrates that the integration of Internet source data, like Google Flu Trends, into surveillance systems can enhance traditional, provider-based networks.

general guideline to target practices with high reporting rates and high numbers of patient visits (per capita) [6]. Polgreen et al. (2009) recently described a computational method for selecting ILINet providers so as to maximize coverage, that is, the number of people living within a specified distance of a provider [17]. They applied the approach to optimizing the placement of the 22 providers in the Iowa ILINet. While their algorithm ensures maximum coverage, it is not clear that maximum coverage is, in general, the most appropriate criterion for building a statistically informative ILINet.

In 2008, Google.org launched Google Flu Trends, a website that translates the daily number of Googles search terms associated with signs, symptoms, and treatment for acute respiratory infections into an estimate of the number of ILI patients per 100,000 people. It was shown that Google Flu Trends reliably estimates national influenza activity in the US [18], the state of Utah [18], and in some European countries [19], but it provided imperfect data regarding the 2009 H1N1 pandemic in New Zealand [20]. We assessed the correlation between Google Flu Trends for Texas and Texas' ILINet data and found a correlation of 0.87, similar to those presented in Ginsberg et al. 2009 [18] (See Text S1). The Google Flu Trends website includes ILI-related search activity down to the level of cities (in beta version as of November 2011). Thus, Google Flu Trends may serve as a valuable resource for influenza detection and forecasting if effectively integrated with public health data such as those coming from state ILINets.

Here, we present an evaluation of the Texas Influenza-Like-Illness Surveillance Network (ILINet), in terms of its ability to forecast statewide hospitalizations due to influenza (ICD9 487 and 488) and unspecified pneumonia (ICD9 486). Although we henceforth refer to this subset of hospitalizations as *influenza-like hospitalizations*, we emphasize that these data do not perfectly reflect influenza-related hospitalizations: some unrelated pneumonias may be classified under ICD9 486, and some influenza cases may not be correctly diagnosed and/or recorded as influenza. Nonetheless, this subset of hospitalizations likely includes a large fraction of hospitalized influenza cases and exhibits strong seasonal dynamics that mirror ILINet trends. The inclusion of all three ICD9 codes was suggested by health officials at Texas DSHS who seek to use ILINet to ascertain seasonal influenza-related hospitalization rates throughout the state (Texas DSHS contract numbers $2009-032591$ and $2011-037903$). Hospitalizations associated with these three codes in Texas accounted for between 20 and 35% of all hospitalizations due to infections and roughly 9.5 billion dollars of hospitalization payments in 2008 (See Text S1).

Using almost a decade of state-level ILINet and hospitalization data, we find that the existing network performs reasonably well in its ability to predict *influenza-like hospitalizations*. However, smaller, more carefully chosen sets of providers should yield higher quality surveillance data, which can be further enhanced with the integration of state-level Google Flu Trends data. For this analysis, we adapted a new, computationally tractable, multilinear regression approach to solving complex subset selection problems. The details of this method are presented below and can be tailored to meet a broad range of surveillance objectives.

Results

Using a submodular ILINet optimization algorithm, we investigate two scenarios for improving the Texas ILINet: designing a network from scratch and augmenting the existing network. We then evaluate the utility of incorporating Google Flu Trends as a virtual provider into an existing ILINet.

Designing a New ILINet

To construct new sentinel surveillance networks, we choose individual providers sequentially from a pool of approximately 2000 mock providers, one for each zip code in Texas, until we reach 200 total providers. At each step, the provider that most improves the quality of the epidemiological information produced by the network is added to the network. We optimize and evaluate the networks in terms of the time-lagged statistical correlation between aggregated ILINet provider reports (simulated by the model) and actual statewide *influenza-like hospitalizations*. Specifically, for each candidate network, we perform a least squares multilinear regression from the simulated ILINet time series to the actual Texas hospitalization time series, and use the coefficient of determination, R^2, as the indicator of ILINet performance. Henceforth, we will refer to these models as *ILINet regression models*.

We compare the networks generated by this method to networks generated by two naive models and a published computational method [17] (Figure 1). *Random* selection models an open call for providers and entails selecting providers randomly with probabilities proportional to their zip code's population; *Greedy* selection prioritizes providers strictly by the population density of their zip code. Submodular optimization significantly outperforms these naive methods, particularly for small networks, with *Random* selection producing slightly more informative networks than *Greedy* selection. The *Geographic* optimization method of Polgreen et al. [17] selects providers to maximize the number of people that live within a specified "coverage distance" of a provider. Submodular optimization consistently produces more informative networks than this method at a 20 mile coverage distance (Figure 1) (5, 10, and 25 mile coverage distances perform worse, not shown). To visualize the relative performance of several of these networks, we compared their estimates of *influenza-like hospitalizations* (by applying each ILINet regression model to simulated ILINet report data) to the true state-wide hospitalization data (Figure 2). The time series estimated by a network designed using submodular optimization more closely and smoothly matches true hospitalizations than both the actual 2008 Texas ILINet and a network designed using geographic optimization (each with 82 providers).

The submodular optimization algorithm is not guaranteed to find the highest performing provider network, and an exhaustive search for the optimal 200 provider network from the pool of 2000 providers is computationally intractable. However, the submod-

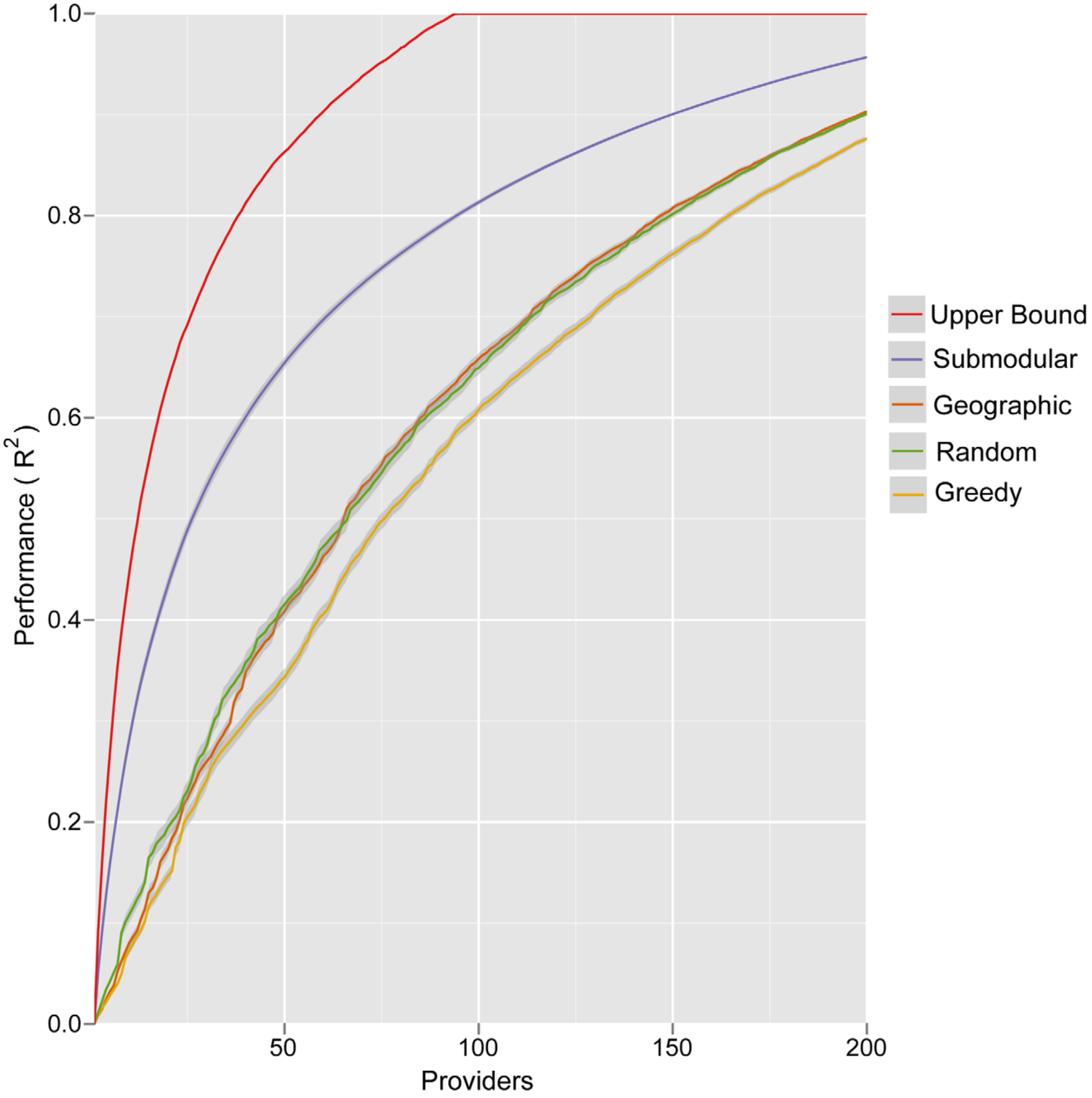

Figure 1. Expected performance of optimized ILINets. Four different methods were used to design Texas ILINets that effectively predict statewide influenza hospitalizations. Submodular optimization (Submodular) outperforms random selection proportional to population density (Random), greedy selection strictly in order of population density (Greedy), and geographic optimization to maximize the number of people that live within 20 miles of a provider [17] (Geographic). The theoretical upper bound for performance (Upper Bound) gives the maximum R^2 possible for a network designed by an exhaustive evaluation of all possible networks of a given size. For each network of each size, the following procedure was repeated 100 times: randomly sample a set of reporting profiles, one for each provider in the network; simulate an ILI time series for each provider in the network; perform an ordinary least squares multilinear regression from the simulated provider reports to the actual statewide influenza hospitalization data. The lines indicate the mean of the resulting R^2 values, and the error bands indicate the middle 90% of resulting R^2 values, reflecting variation stemming from inconsistent provider reporting and informational noise.

doi:10.1371/journal.pcbi.1002472.g001

ular property of the objective function allows us to compute an upper bound on the performance of the optimal network, without knowing its actual composition (Figure 1). The performance gap between the theoretical upper bound and the optimized networks may indicate that the upper bound is loose (higher than the performance of the true optimal network) and/or the existence of better networks that might be found using more powerful optimization methods.

The networks selected by submodular optimization reveal some unexpected design principles. Most of the Texas population resides in Houston and the "I-35 corridor" – a North-South transportation corridor spanning San Antonio, Austin, and Dallas (Figure 3a). The first ten provider locations selected by submodular optimization are spread throughout the eastern half of the state (Figure 4a, pink circles). While most of the providers are concentrated closer to Texas' population belt, only two are

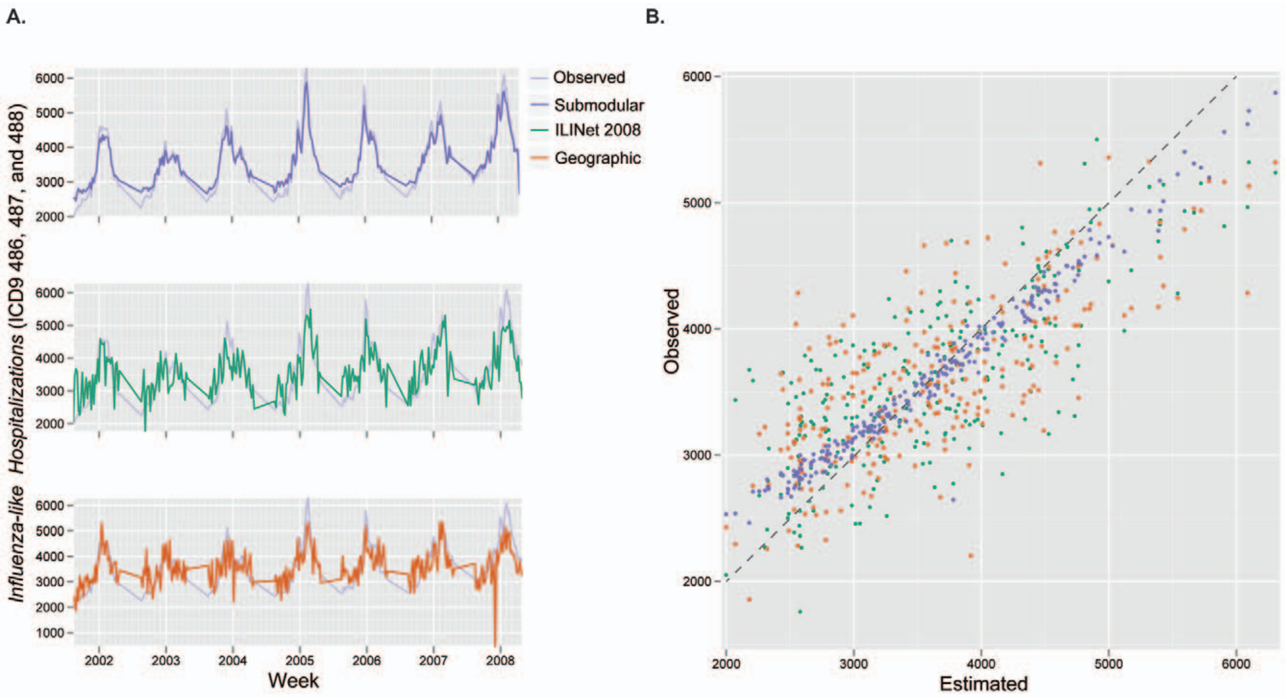

Figure 2. Comparing ILINet estimates to actual state-wide influenza hospitalizations. Statewide hospitalizations are estimated using data from three ILINets: the 2008 Texas ILINet (ILINet 2008), which consisted of 82 providers, and ILINets of the same size that were designed using submodular optimization (Submodular) and maximum coverage optimization with a 20 mile coverage distance (Geographic). (a) The estimates from each network are compared to actual Texas state-wide influenza hospital discharges from 2001–2008 (Observed). (b) The submodular ILINet yields estimates that are consistently closer to observed values than the other two ILINets. For each of the three networks, the following procedure was repeated 100 times: randomly sample a set of reporting profiles, one for each provider in the network; simulate an ILI time series for each provider in the network; perform an ordinary least squares multilinear regression from the simulated provider reports to the actual Texas influenza hospitalization data; and apply resulting regression model to the simulated provider time series data to produce estimates of statewide hospitalizations. The figures are based on averages across the 100 estimated hospitalization time series for each ILINet.
doi:10.1371/journal.pcbi.1002472.g002

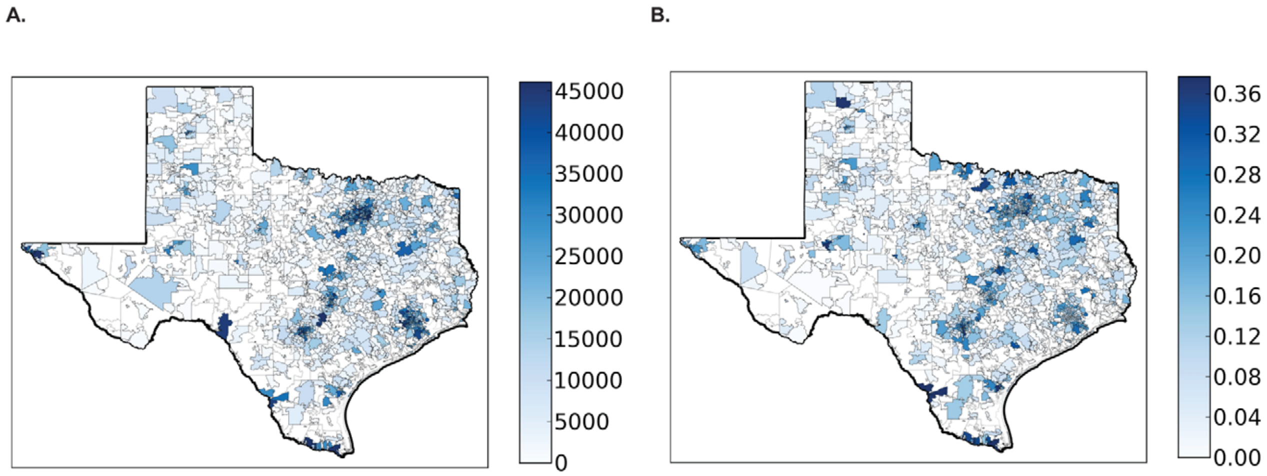

Figure 3. Statewide influenza activity mirrors population distribution. (a) Shading indicates zip code level population sizes, as reported in the 2000 census. (b) Major populations centers exhibit covariation in influenza activity. We performed a principal component analysis (PCA) on the centered hospitalization time series of all zip codes and calculated the time series of the first principal component. Zip codes are shaded according to the R^2 obtained from a regression of the first principal component time series to the influenza hospitalization time series for the zip code. Dark shading indicates high synchrony between influenza activity in the zip code and the first principal component. The correspondence between darkly shaded zip codes in (a) and (b) results from the high degree of synchrony in influenza activity between highly populated zip codes in Texas.
doi:10.1371/journal.pcbi.1002472.g003

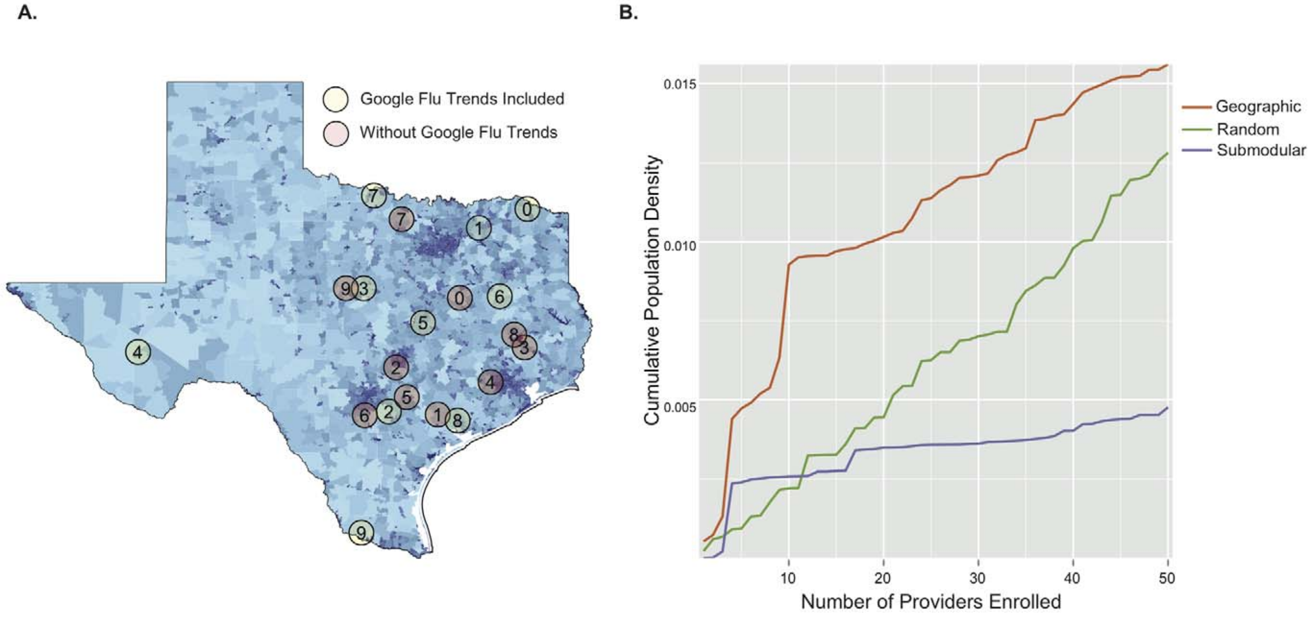

Figure 4. Location and population coverage of optimized ILINets. (a) Shading indicates zip code level population sizes, as reported in the 2000 census. Circles indicate the location (zip code) of the first ten providers selected when Google Flu Trends is included as a provider (green) and when it is not (pink). Numbers indicate selection order, with zero being the first provider selected and nine the tenth provider selected. (b) The cumulative population densities covered increase as each ILINet grows. Cumulative density is estimated by dividing total population of all provider zip codes by total area of all provider zip codes. While ILINets designed using the geographic (orange) and random (green) methods primarily target zip codes with high population densities, submodular optimization (purple) targets zip codes that provide maximal information, regardless of population density. All three networks cover approximately the same total number of people.
doi:10.1371/journal.pcbi.1002472.g004

actually located within Texas' major population centers (in this case, College Station).

The submodular networks are qualitatively different from the networks created by the other algorithms considered, which focus providers within the major population centers (Figure 4b). The higher performance of the submodular ILINets suggest that over-concentration of providers in major population centers is unnecessary. Influenza levels in the major population centers are strongly correlated (Figure 3b). Thus, ILINet information from San Antonio, for example, will also be indicative of influenza levels in Austin and Dallas. This synchrony probably arises, in part, from extensive travel between the major Texas population centers.

Subsampling and Augmenting an ILINet

Using submodular optimization, we augment the 2008 Texas ILINet by first subsampling from the 82 enrolled providers and then adding up to 40 new providers. When subsampling, performance does not reach a maximum until all 82 providers are included in the network (Figure 5), indicating that each provider adds predictive value to the network. However, the theoretical upper bound plateaus around 40 providers, suggesting that smaller (more optimally chosen) networks of equal predictive value may exist. During the second stage, 40 additional providers improve the R^2 objective by 33%. Most of these providers are located in relatively remote areas of the state.

We also considered inclusion of Internet trend data sources as virtual providers, specifically, the freely available Google Flu Trends data for the state of Texas [21]. Google Flu Trends alone is able to explain about 60% of the variation in state-wide hospitalizations; it outperforms the 2008 Texas ILINet and matches the performance of a network with 44 traditional providers constructed from scratch using submodular optimization (Figure 6). However, the best networks include both traditional providers and Google Flu Trends. For example, by adding 50 providers to Google Flu Trends using submodular optimization, we improve the R^2 objective by a third and halve the optimality gap (from a trivial upper bound of one). The additional providers are located in non-urban areas (Figure 4a, green circles) distinct from those selected when Google Flu Trends is not allowed as a provider.

Out-of-Sample Validation

To further validate our methodology, we simulated the real-world scenario in which historical data are used to design an ILINet and build forecasting models, and then current ILINet reports are used to make forecasts. Specifically, we used $2001-2007$ data to design ILINets and estimate multilinear regression models relating *influenza-like hospitalizations* to mock provider reports, and then used 2008 data to test the models' ability to forecast *influenza-like hospitalizations*. For networks with fewer than 150 providers, the ILINets designed using submodular optimization consistently outperform ILINets designed using the other three strategies (Figure 7). Above 100 providers, the predictive performance of the submodular optimization ILINet begins to decline with additional providers. As the number of providers approaches 222 (the number of weeks in the training period), the estimated prediction models become overfit to the $2001-2007$ period. Thus, the slightly increased performance of the *Random* method over the submodular optimization after 175 providers is spurious. For the $\hat{R}^2$ values presented in Figure 7, the effect of noise and variable reporting are integrated out when calculating the expected provider reports. An alternative approach to out-of-sample validation is presented in Text S1; it yields the same rank-order of model performance.

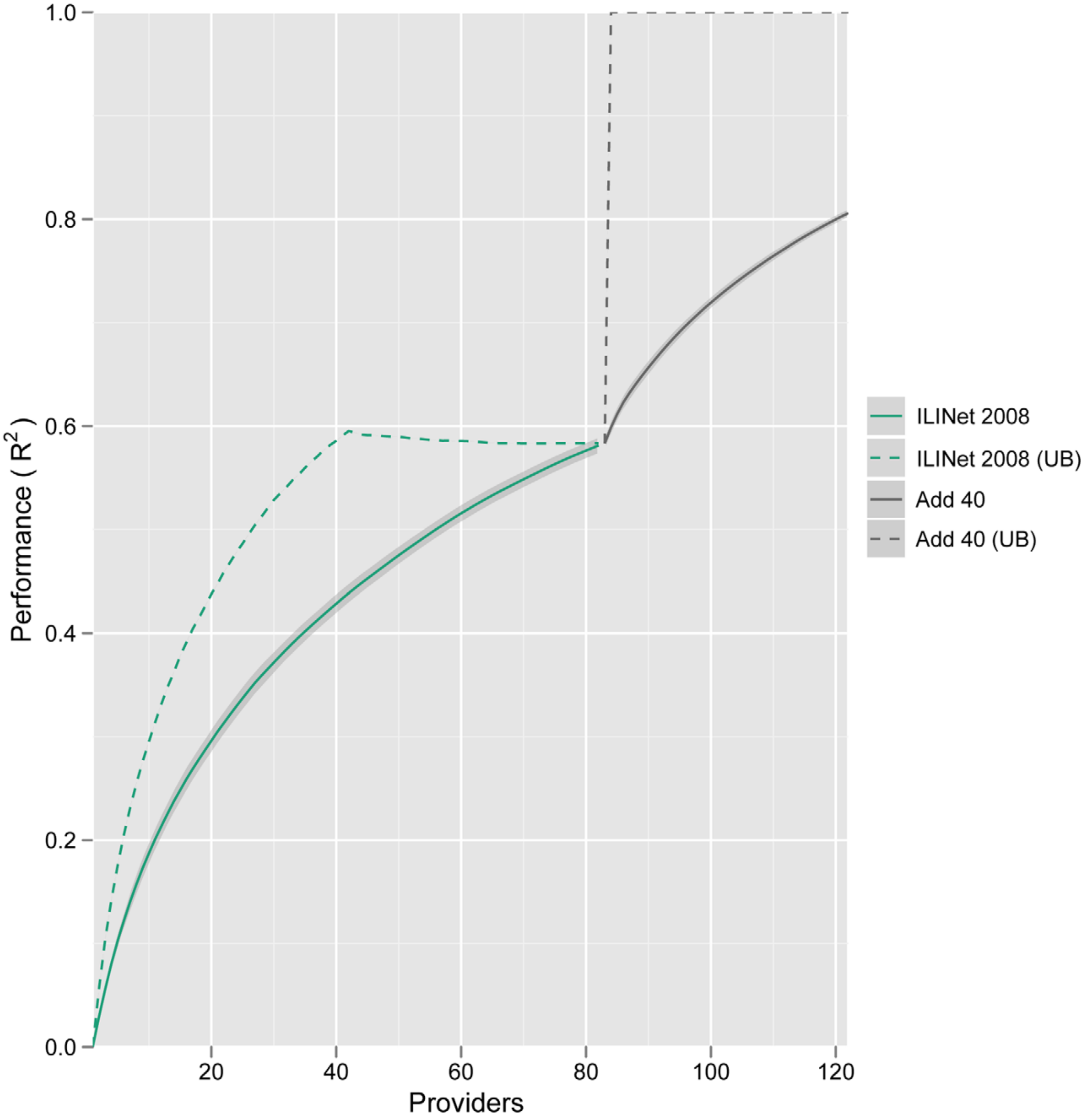

Figure 5. Augmenting an existing ILINet. This compares theoretical upper bounds (dashed lines) to the performance of a submodular optimized ILINet built by first subsampling the 82 zip codes of providers actually enrolled in Texas' 2008 ILINet (green) and then adding 40 additional providers from elsewhere in the state (gray). The error bands indicate the middle 90% of resulting R^2 values, and reflect variation stemming from inconsistent provider reporting rates and informational noise.
doi:10.1371/journal.pcbi.1002472.g005

Discussion

Since the mid twentieth century, influenza surveillance has been recognized as an increasingly complex problem of global concern [22]. However, the majority of statistical research has focused on the analysis of surveillance data rather than the data collection itself, with a few notable exceptions [12,17]. High quality data is essential for effectively monitoring seasonal dynamics, detecting anomalies, such as emerging pandemic strains, and implementing effective time-sensitive control measures. Using a new method for optimizing provider-based surveillance systems, we have shown that the Texas state ILINet would benefit from the inclusion of a few strategically selected providers and the use of Internet data streams.

Our method works by iteratively selecting providers that contribute the most information about *influenza-like hospitalizations*. We quantified the performance of various ILINets using the coefficient of determination (R^2) resulting from a multi-linear regression between each provider's time series and state-wide *influenza-like hospitalizations*. Importantly, these simulated providers have reporting rates and error distributions estimated from actual ILINet providers in Texas (see Text S1). The result is a prioritized list of zip codes for inclusion in an ILINet that can be used for future ILINet recruiting. Although this analysis was specifically

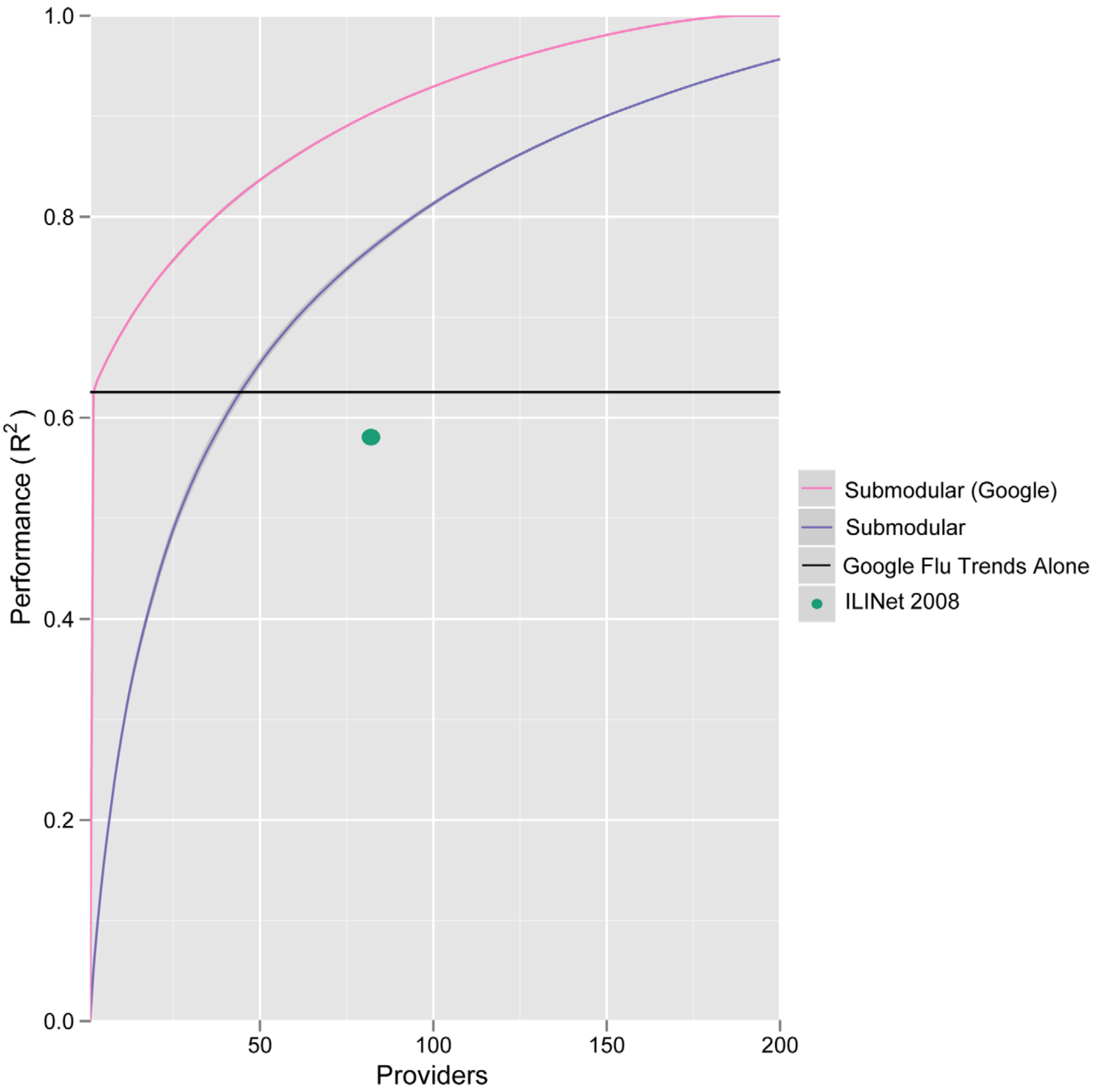

Figure 6. Google Flu Trends as a virtual ILINet provider. When state-level Google Flu Trends is treated as a possible provider, submodular optimization choses it as the first (most informative) provider for the Texas ILINet, and results in a high performing network (pink line). Alone (black line), the Google Flu Trends provider performs as well as a traditional submodular optimized network (blue line) containing 44 providers (intersection of black and purple lines) and outperforms the actual 2008 Texas ILINet (green dot).
doi:10.1371/journal.pcbi.1002472.g006

motivated by the Texas DSHS interest in predicting hospitalizations with ICD9 codes 486, 487, and 488, our method can be readily extended to design a network for any disease or influenza definition with the appropriate historical data. In general, the method requires both historical provider reports and historical time series of the prediction target. However, if one has reasonable estimates of provider reporting rates and informational noise from another source (e.g., estimates from a surveillance network in another region or for another disease), then historical provider reports are not necessary.

ILINet provider reports do not necessarily reflect true influenza activity. Rather they are supposed to indicate the number of patients that meet the clinical ILI case definition, which results in a substantial number of false positives (reported non-influenza cases) and false negatives (missed cases of influenza) [23]. The case definition for ILI is often loosely applied, further confounding the relationship between these measures and true influenza. Similarly, the ICD9 codes used in our analysis do not correspond perfectly to influenza hospitalizations: some influenza cases will fail to be classified under those codes, and some non-influenza cases will be. Nonetheless, public health agencies are interested in monitoring and forecasting the large numbers of costly hospitalizations associated with these codes. We find that ILINet surveillance data correlates strongly with this set of

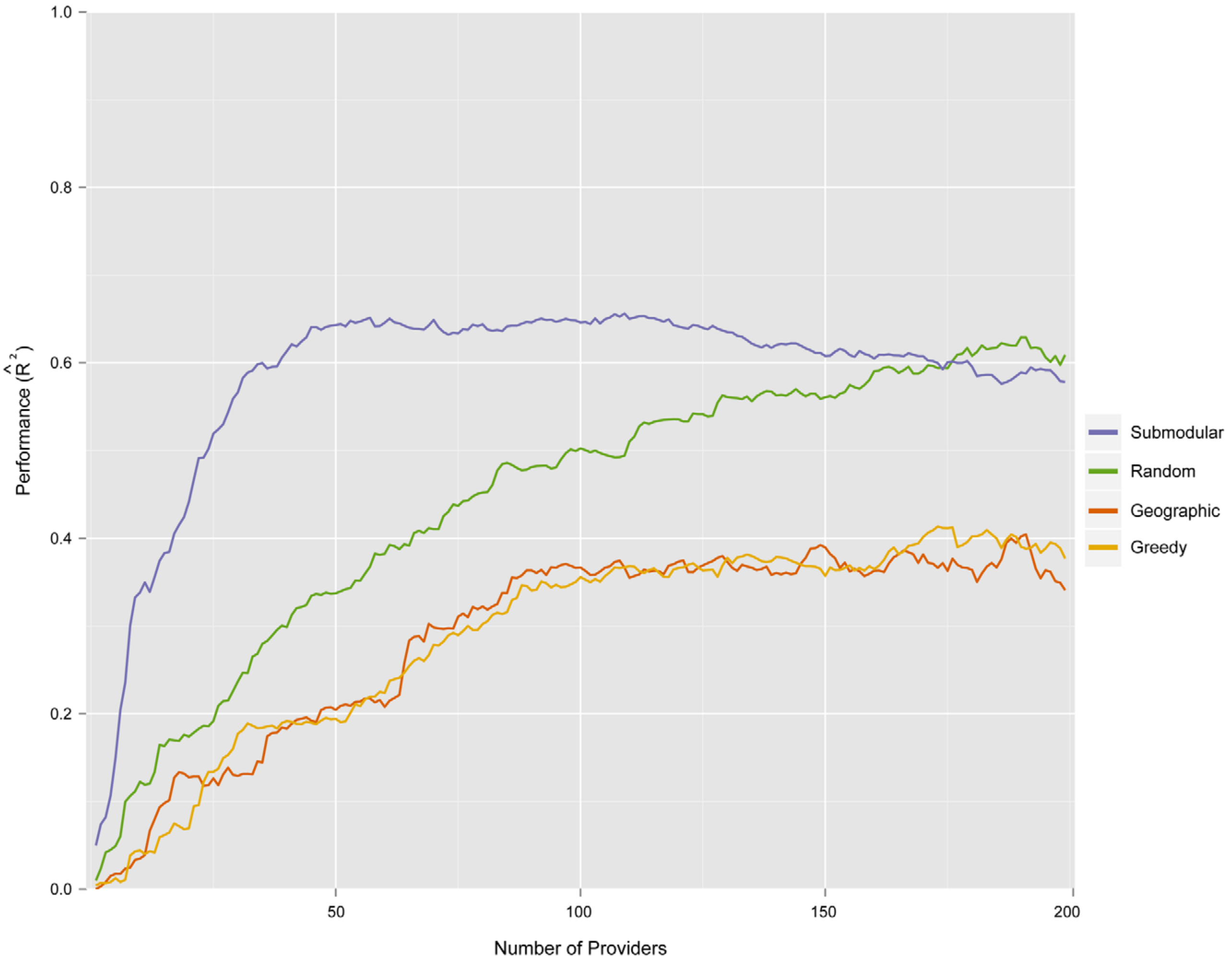

Figure 7. Predictive performance of ILINets. Data from the 2001–2007 period were used to design ILINets and estimate multilinear regression prediction models. The predictive performance of the ILINets (y-axis) is based on a comparison between the models' predictions for 2008 hospitalizations (from mock provider reports) and actual 2008 hospitalization data. For almost all network sizes, Submodular optimization (Submodular) outperforms random selection proportional to population density (Random), greedy selection strictly in order of population density (Greedy), and geographic optimization to maximize the number of people that live within 20 miles of a provider [17] (Geographic). The leveling-off of performance around 100 providers is likely a result of over-fitting, given that there were only 222 historical time-points used to estimate the original model.

doi:10.1371/journal.pcbi.1002472.g007

influenza-like hospitalizations, and that the networks can be designed to be even more informative.

Although we provide only a single example here, this optimization method can be readily applied to designing surveillance networks for a wide range of diseases on any geographic scale, provided historical data are available and the goals of the surveillance network can be quantified. For example, surveillance networks could be designed to detect emerging strains of influenza on a global scale, monitor influenza in countries without surveillance networks, or track other infectious diseases such as malaria, whooping cough, or tuberculosis or non-infectious diseases and chronic conditions such as asthma, diabetes, cancer or obesity that exhibit heterogeneity in space, time or by population subgroup. As we have shown with Google Flu Trends, our method can be leveraged to evaluate the potential utility of incorporating other Internet trend data mined from search, social media, and online commerce platforms into traditional surveillance systems.

While optimized networks meet their specified goals, they may suffer from over optimization and be unable to provide valuable information for other diseases or even for the focal disease during atypical situations. For example, a surveillance network designed for detecting the early emergence of pandemic influenza may look very different from one optimized to monitor seasonal influenza. Furthermore, an ILINet optimized to predict *influenza-like hospitalizations* in a specific socio-economic group, geographic region, or race/ethnicity may look very different from an ILINet optimized to predict state-wide hospitalizations. When optimizing networks, it is thus important to carefully consider the full range of possible applications of the network and integrate diverse objectives into the optimization analysis.

The optimized Texas ILINets described above exhibit much less redundancy (geographic overlap in providers) than the actual Texas ILINet. Whereas CDC guidelines have led Texas DSHS to focus the majority of recruitment on high population centers, the optimizer only sparsely covered the major urban areas because of their synchrony in influenza activity. This is an important distinction between submodular optimization and the other methods considered (*Geographic*, *Random* and *Greedy*). The submodular method does not track population density and instead adds providers who contribute the most marginal information to the network. Consequently, it places far more providers in rural areas than the other methods (Figure 4b). There can be substantial year-to-year variation in spatial synchrony for seasonal influenza, driven by the predominant influenza strains and commuter traffic between population centers [24]. As long as the historical data used during optimization reflect this stochasticity, the resulting networks will be robust. However, synchrony by geography and population density does not occur for all diseases including emerging pandemic influenza [24]; thus the relatively sparse networks designed for forecasting seasonal influenza hospitalizations may not be appropriate for other surveillance objectives, like detecting emerging pandemic strains or other rare events. For example, a recent study of influenza surveillance in Beijing, PRC suggested that large hospitals provided the best surveillance information for seasonal influenza, while smaller provincial hospitals were more useful for monitoring H5N1 [12].

Although our method outperforms the *Maximal Coverage Method* (MCM), referred to as *Geographic*, proposed by Polgreen et al. (2009), there are several caveats. First, population densities and travel patterns within Texas are highly non-uniform. The two methods might perform similarly for regions with greater spatial uniformity. Second, our method is data intensive, requiring historical surveillance data that may not be available, for example, in developing nations, whereas the population density data required for MCM is widely available. However, the type of data used in this study is readily available to most state public health agencies in the United States. For example, the CDC's Influenza Hospitalization Network (FluSurv-NET) collects weekly reports on laboratory confirmed influenza-related hospitalizations in fourteen states. In addition, alternative internet-based data sources like Google Flu Trends are becoming available. Third, as discussed above, our networks are optimized towards specific goals and may thus have no expected level of performance for alternate surveillance goals. Important future research should focus on designing networks able to perform well under a range of surveillance goals. Fourth, neither ILINet data nor *influenza-like hospitalizations* correspond perfectly to actual influenza activity. One could instead optimize ILINets using historical time series of laboratory-confirmed cases of influenza. Although some provider locations and the estimated regression models may change, we conjecture that the general geospatial distribution of providers will not change significantly. Fourth, we followed Polgreen et al. (2009)'s use of Euclidean distances. However, travel distance is known to correlate more strongly with influenza transmission than Euclidean distance [24], and thus alternative distance metrics might improve the performance of the MCM method. Finally, while submodular optimization generally outperforms the other design methods in out-of-sample prediction of *influenza-like hospitalizations*, it suffers from overfitting when the number of providers in the network approaches the number of data points in the historical time series.

The impressive performance of Google Flu Trends leads us to question the role of traditional methods, such as provider-based surveillance networks, in next generation disease surveillance systems. While Texas Google Flu Trends alone providers almost as much information about state-wide influenza hospital discharges as the entire 2008 Texas ILINet, an optimized ILINet of the same size contains 33% more information than Google Flu Trends alone. Adding Google Flu Trends to this optimized network as a virtual provider increases its performance by an additional 12.5%. Internet driven data streams, like Google Flu Trends, may have age and socio-economic biases that over-represent certain groups, a possible explanation for the difference in providers selected when Google Flu Trends is included, Figure 4a. Given the relatively low cost of voluntary provider surveillance networks, synergistic approaches that combine data from conventional and Internet sources offer a promising path forward for public health surveillance.

This optimization method was conceived through a collaboration between The University of Texas at Austin and the Texas Department of State Health Services to evaluate and improve the Texas ILINet. The development and utility of quantitative methods to support public health decision making hinges on the continued partnership between researchers and public health agencies.

Materials and Methods

Data

The Texas Department of State Health Services (DSHS) provided (1) ILINet data containing weekly records from $2001-2010$ reporting the number of patients with influenza-like-illness and the total number of patients seen by each provider in the network, and (2) individual discharge records for every hospital in Texas from $2001-2007$ (excluding hospitals in counties with less than 35,000 inhabitants, in counties with less than 100 total hospital beds, or those hospitals that do not seek insurance payment or government reimbursement). We classified all hospital discharges containing ICD9 codes of 486, 487, or 488 as influenza-related. Google Flu Trends data was downloaded from the Google Flu Trends site [21] and contains estimates of ILI cases per 100,000 physician visits determined using Google searches [25]. Data on population size and density was obtained from the 2000 census [26].

Provider Reporting Model

The first step in the ILINet optimization is to build a data-driven model reflecting actual provider reporting rates and informational noise, that is, inconsistencies between provider reports and true local influenza prevalence.

We model reporting as a Markov process, where each provider is in a "reporting" or "non-reporting" state. A provider in the reporting state enters weekly reports, while a provider in the non-reporting state does not enter reports. At the end of each week, providers independently transition between the reporting and non-reporting states. Such a Markov process model allows for streaks of reporting and streaks of non-reporting for each provider, which is typical for ILINet providers. We estimate transition probabilities between states from actual ILINet provider report data. For each provider, the transition probability from reporting to non-reporting is estimated by dividing the number of times the transition occurred by the number of times any transition out of reporting is observed. The probabilities of remaining in the current reporting state and transitioning from non-reporting to reporting are estimated similarly.

We model noise in reports using a standard regression noise model of the form

$$\mathrm{Provider-report(i)} = c_0 + c_1 \mathrm{Percent-ILI(i)} + N(0,\sigma^2), \quad (1)$$

where $\mathrm{Provider-report}(i)$ denotes the number of ILI cases reported by the provider in week i; $\mathrm{Percent-ILI}(i)$ denotes the estimated prevalence of ILI in the provider's zip code in week i; c_0

and c_1 are regression constants fixed for the provider; and $N(0,\sigma^2)$ is a normally distributed noise term with variance σ^2 also fixed for the provider. For existing providers, we use empirical time series (their past ILINet reporting data matched with local ILI prevalence, described below) to estimate the constants c_0,c_1, and σ^2 using least squares linear regression. This noise model has the intuitive interpretation that each provider's reports are a noisy reading of the percent of the population with ILI in the provider's zip code.

We use the Texas hospital discharge data to estimate the local ILI prevalences ($\mathrm{Percent}-\mathrm{ILI}(i)$) for each zip code. Given an estimate of the influenza hospitalization rate [27] and assuming that each individual with ILI is hospitalized independently, we can obtain a distribution for the number of influenza-related hospitalizations in a zip code, given the number of ILI cases in the zip code. Using Bayes rule, a uniform prior, and the real number of influenza-related hospitalizations (from the hospital discharge data), we derive distributions for the number of ILI cases for each zip code and each week. We then set $\mathrm{Percent}-\mathrm{ILI(i)}$ for each zip code equal to the mean of the distribution of ILI cases in that zip code for week i, divided by the population of the zip code.

Generating Pools of Mock Providers

The second step in the ILINet optimization is to generate a pool of mock providers. For each actual provider in the Texas ILINet, we estimate a reporting profile specified by [1)] transition probabilities between reporting and non-reporting (Markov) states, and the constants c_0,c_1 and σ^2, modeling noise in the weekly ILI reports. To generate a mock provider in a specified zip code, we select a uniformly random reporting profile out of all reporting profiles estimated from existing providers. The generated mock providers are thereby given reporting characteristics typical of existing providers. We can then generate an ILI report time series for a mock provider, by 1) generating reports only during reporting weeks, and calculating reports using equation (1) with the constants given in the provider's reporting profile and estimates of $\mathrm{Percent}-\mathrm{ILI}(i)$ for the mock provider's zip code.

We select providers from pools consisting of a single mock provider from each zip code. Zip codes offer a convenient spatial resolution, because they have geographic specificity and are recorded in both the Texas ILINet and hospital discharge data. The optimization algorithm is not aware of a mock provider's reporting profile when the provider is selected (discussed below).

Provider Selection Optimization

The final step in our ILINet design method is selecting an optimized subset of providers from the mock provider pool. We seek the subset that most effectively predicts a target time series (henceforth, goal), as measured by the coefficient of determination (R^2) from a least squares multilinear regression to the goal from the report time series for all providers in the subset. Specifically, the objective function is given by

$$R^2(\mathbf{G},S)=\frac{\mathrm{Var}(\mathbf{G})-\mathrm{Var}(\mathbf{G}-\sum_{i\in S}\alpha_i\mathbf{P_i})}{\mathrm{Var}(\mathbf{G})},$$

where $\mathbf{G}$ is the goal random variable; S is a subset of the mock provider pool; $\mathbf{P_i}$ are provider reports for provider i; and the α_i are the best multilinear regression coefficients (values that minimize the second term in the numerator).

There are several advantages to this objective function. First, it allows us to optimize an ILINet for predicting a particular random variable. Here, we set the goal to be state-wide influenza-related hospitalizations for Texas. This method can be applied similarly to design surveillance networks that predict, for example, morbidity and/or mortality within specific age groups or high risk groups.

Second, the objective function is submodular in the set of providers, S [28], implying generally that adding a new provider to a small network will improve performance more than adding the provider to a larger network. The submodular property enables computationally efficient searches for near optimal networks and guarantees a good level of performance from the resulting network [29]. Without a submodular objective function, optimization of a k provider ILINet may require an exhaustive search of all subsets of k providers from the provider pool, which quickly becomes intractable. For example, an exhaustive search for the optimal 200 provider Texas ILINet from our pool of approximately 2000 mock providers would require roughly 10^{660} regressions.

Taking advantage of the submodular property, we rapidly build high performing networks (with k providers) according to the following algorithm:

1. Let P be the entire provider pool, S be the providers selected thus far, and $f(S)$ be a submodular function in S. We begin without any providers in S.
2. Repeat until there are k providers in S:

(a) Let x be the provider in $P-S$ that maximizes $f(S+x)-f(S)$

(b) Add x to S.

This is guaranteed to produce a network that performs within a fraction of $1-\frac{1}{e}$ of the optimal network [28]. The submodularity property also allows us to compute a posterior bound on the distance from optimality, which is often much better than $1-\frac{1}{e}$. Finally, even if implemented naively, the algorithm only requires approximately $10^{5.6}$ regressions to select 200 providers from a pool of 2000.

When optimizing, it is important to consider potential noise (underreporting and discrepancies between provider reports and actual ILI activity in the zip code). However, we assume that one cannot predict the performance of a particular provider before the provider is recruited into the network. To address this issue, the optimization's objective function is an expectation over the possible provider reporting profiles. Specifically, we define $\tilde{\xi}$ as a random variable describing the provider reporting profile for the entire pool of mock providers. If $\hat{\xi}$ is a specific reporting profile, then the R^2 objective function can be written as

$$R^2(\mathbf{G},S,\hat{\xi})=\frac{\mathrm{Var}(\mathbf{G})-\mathrm{Var}(\mathbf{G}-\sum_{i\in S}\alpha_i\mathbf{P_i}(\hat{\xi}))}{\mathrm{Var}(\mathbf{G})}.$$

To design the ILINet, we solve the following optimization problem

$$\max_{S\subseteq P}E_{\tilde{\xi}}[R^2(\mathbf{G},S,\tilde{\xi})].$$

The objective function is a convex combination of submodular functions, and thus is also submodular. This allows us to use the above algorithm along with its theoretical guarantees to design ILINets using a realistic model of reporting practices and informational stochasticity, without assuming that the designer knows the quality of specific providers *a priori*.

Maximal Coverage Model

We implemented the *Maximal coverage model* (MCM) following Polgreen et al. (2009). Briefly, a greedy algorithm was used to minimize the number of people in Texas who live outside a predefined coverage distance, C, of at least one provider in the selected set, S. A general version of this algorithm was developed by Church and Re Velle (1974) to solve this class of MCM's [30]. As per Polgreen et al. (2009), we assumed that the population density of each zip code exists entirely at the geographic center of the zip code and used Euclidean distance to measure the distance between zip codes. Using a matrix of inter-zip code distances we select providers iteratively, choosing zip codes that cover the greatest amount of population density within the pre-defined coverage distance, C. We considered $C = 5$, 10, 20, and 25 miles, and found that $C = 20$ miles yielded the most informative networks.

Naive Methods

We used two naive methods to model common design practices for state-level provider-based surveillance networks.

1. *Greedy selection by population density* - All zip codes were ordered by population density and added to the provider pool P. Providers are then moved from P to the selected set S from highest to lowest density. The algorithm stops when S reaches a predetermined size or P is empty.
2. *Uniform random by population size* - Zip codes are randomly selected from P and moved to S with a probability proportional to their population size. The algorithm proceeds until either S reaches a pre-determined size or P is empty.

Principal Component Analysis of Hospitalization Time Series across Texas Zip Codes

To analyze similarities in ILI hospitalizations across different zip codes, we apply principal component analysis (PCA) [31]. Specifically, we perform PCA on the centered (mean zero), standardized (unit variance) hospitalization time series of all zip codes in Texas. We first compute a time series for the first principal component, and then compute an R^2 for each zip code, based on a linear regression from the first principal component to the zip code's centered, standardized hospitalizations. Zip codes with high R^2 values have hospitalization patterns that exhibit high temporal synchronicity with the first principal component.

Out-of-Sample Validation

To validate our method, we first use submodular optimization to create a provider network of 200 providers, using only data from 2001 to 2007, and then evaluate the performance of the network in predicting 2008 influenza-like hospitalizations. Specifically, after creating the 200-provider network (S_{train}), we use actual hospitalization data and mock provider reports for the 2001–2007 period to fit a multilinear regression model of the form $G^{\text{train}}(t) = \sum_{i \in S_{\text{train}}} \alpha_i^{\text{train}} \mathbf{P}_\mathbf{i}^{\text{train}}(t-2)$ where $G^{\text{train}}(t)$ is time series of state-wide influenza-like hospitalizations at week t for weeks in 2001 to 2007, $P_i^{\text{train}}(t-2)$ is the mock report time series of provider i during week $t-2$ for weeks in 2001 to 2007, and α_i^{train} is the best multilinear regression coefficient associated with provider i.

We then use the estimated multilinear regression function to forecast state-wide influenza-like hospitalization during 2008 from mock provider reports of 2008, and compare these forecasts to actual 2008 hospitalization data. This simulates a real-world prediction, where only historical data is available to create the provider network (S_{train}) and estimate the prediction function (α_i^{train}'s), and then the most recent provider reports ($\mathbf{P}_\mathbf{i}^{2008}$'s) are used to make predictions. We evaluate the 2008 predictions using a variance reduction measure similar to R^2, except that the multilinear prediction model uses coefficients estimated from prior data, as given by

$$\hat{R}^2(\mathbf{G^{2008}}, S^{\text{train}}) = \frac{\mathrm{Var}(\mathbf{G^{2008}}) - \mathrm{Var}(\mathbf{G^{2008}} - \sum_{i \in S_{\text{train}}} \alpha_i^{\text{train}} \cdot E_\xi[P_i^{2008}(\xi)])}{\mathrm{Var}(\mathbf{G^{2008}})},$$

where G^{2008} is the hospitalization time series in 2008, ξ is the provider noise profile, and $P_i^{2008}(\xi)$ are the mock provider reports in 2008. Importantly, we first calculate an expected value for the provider reports, $P_i^{2008}(\xi)$, given the noise profiles ξ, before calculating $\hat{R}^2$. We also considered an alternative validation method in which we first calculate an R^2 for each provider report and noise-profile combination, and then analyze the resulting distribution of R^2 values (see Text S1 for results).

Supporting Information

Figure S1 Proportion of hospitalizations associated with ICD9s 486, 487 and 488 - We present the proportion of respiratory illness related hospitalizations that were also associated with ICD9s 486, 487 and 488. The total number of respiratory illness related hospitalizations were estimated from the Texas hospitalization database, the same database used to determine the number of ICD9 486, 487 and 488 associated cases. There is a strong seasonality in the proportion, with peaks in the winter between 0.30 and 0.37 and valleys in the summer around 0.24.
(TIF)

Figure S2 Weekly costs associated with ICD9s 486, 487 and 488 - The total weekly billing charges associated with *influenza-like hospitalizations* are plotted from the end of 2001 through the beginning of 2009. On average 500 million dollars of hospital charges were billed per month to patients associated with ICD9s 486, 487 and 488. However, it is important to note the over two-fold increase in this amount since 2002. For the 2007–2008 influenza season this increase corresponded to a total billed amount of 9.3 billion dollars. This represents nearly 1 percent of the yearly GDP in Texas, which is not much less than the year-to-year economic growth.
(TIF)

Figure S3 Texas ILINet provider reporting rates - (a) Histograms are presented for the four transition probabilities used in our Markov model of provider reporting. The change in skew between panels *i* and *iv* as compared to panels *ii* and *iii* is expected given the observation of "streaky" reporting of ILINet providers in Texas. The providers with a score of one in panel *ii* are those ideal providers who are likely to resume reporting after missing a week. (b) A scatter plot of the values in S3a- *i* and S3a- *ii*, Report given Reported and Report given Failed to Report, are presented to indicate that there are both reliable and unreliable providers enrolled in the Texas ILINet, with darker blue indicating a more reliable provider and light-blue to white a less reliable provider.
(TIF)

Figure S4 Out-of-Sample Model Validation - We used data from 2001–2007 to design ILINets and to fit multi-linear prediction functions, and then generated provider-report based forecasts of hospitalizations during 2008 (without using any data from 2008) and compared these predictions to actual 2008 hospitalization data (see text for details). The $\tilde{R}^2$ values reflect the

predictive performance of the different ILINets. For each ILINet, we predicted 100 time series from simulated provider reports, each time drawing random deviates from the provider noise and reporting distributions, and then compared them to actual 2008 hospitalizations by calculating $\tilde{R}^2$. Lines indicate the average $\tilde{R}^2$ and shaded regions indicate the middle 90% of the $\tilde{R}^2$ distribution. Negative values indicate that the predicted hospitalization time series are more variable than the actual time series. The increasingly poor performance and uncertainty with additional providers is a result of over-fitting of the prediction model to data from the 2001–2007 training period. The submodular method is the only one to yield ILINets with a $\tilde{R}^2$ greater than zero.
(TIF)

Figure S5 The importance of realistic reporting rates and noise - We compared the first ten providers selected by the submodular optimization method when providers either contained (a) perfect information and perfect reporting rates or (b) were subject to the patterns of imperfect and variable reporting exhibited by actual ILINet providers. When simulated providers had reporting probabilities and noise similar to actual providers the resulting network contained more geographic redundancy than one built from simulated providers with perfect information and reporting rates. All results presented in the manuscript were determined using simulated providers with patterns of imperfect and variable reporting derived from actual ILINet data. The stark difference highlights the importance of incorporating the characteristics of actual ILINet provider reporting.
(TIF)

Text S1 In text S1, we present the results of five supplementary analyses. 1) The importance of *influenza-like hospitalizations* in terms of total respiratory disease related hospitalizations and health care charges in Texas, 2) The details of actual ILINet provider reporting in Texas, the data described here were used to derive our provider reporting model, 3) The time-lagged, linear relationship between *influenza-like hospitalizations*, ILINet, and Google Flu Trends in Texas, 4) Additional model validation results, which support and confirm those presented in the main text, and 5) The importance of incorporating realistic provider reporting rates and noise illustrated by the dramatic difference in the results when perfect information and reporting is assumed.
(PDF)

Acknowledgments

The authors thank Marilyn Felkner and Jeff Taylor at the Texas Department of State Health Services and the Meyers and Wilke labs at the University of Texas at Austin for useful comments and discussion.

Author Contributions

Conceived and designed the experiments: SVS NBD LAM. Performed the experiments: SVS NBD. Analyzed the data: SVS NBD. Contributed reagents/materials/analysis tools: SVS NBD. Wrote the paper: SVS NBD LAM.

References

1. Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med 5: e151.
2. Khan AS, Fleischauer A, Casani J, Groseclose SL (2010) The Next Public Health Revolution: Public Health Information Fusion and Social Networks. Am J Public Health 100: 1237–1242.
3. Mnatsakanyan ZR, Burkom HS, Coberly JS, Lombardo JS (2011) Bayesian Information Fusion Networks for Biosurveillance Applications. J Am Med Inform Assoc 16(6): 855–863.
4. Bush GW (2007) Homeland security presidential directive 21: public health and medical preparedness. Available: http://www.dhs.gov/xabout/laws/gc_1219263961449.shtm. Accessed February 13th 2012.
5. US Centers for Disease Control and Prevention (2010) National biosurveillance strategy for human health. Available: http://www.cdc.gov/osels/ph_surveillance/bc.html. Accessed February 13th 2012.
6. Clothier H, Turner J, Hampson A, Kelly H (2006) Geographic representativeness for sentinel influenza surveillance: implications for routine surveillance and pandemic preparedness. Aust NZ J Public Health 30: 337–341.
7. Deckers JG, Paget WJ, Schellevis FG, Fleming DM (2006) European primary care surveillance networks: their structure and operation. Fam Pract 23: 151–158.
8. Carrat F, Flahault A, Boussard E, Farran N, Dangoumau L, et al. (1998) Surveillance of influenza-like illness in france. The example of the 1995/1996 epidemic. J Epidemiol Community Health 52: 32S–38S.
9. Ordobás M, Zorrilla B, Arias P (1995) Influenza in madrid, spain, 1991–92: validity of the sentinel network. J Epidemiol Community Health 49: 14–16.
10. Viboud C, Boëlle P, Carrat F, Valleron A, Flahault A (2003) Prediction of the spread of influenza epidemics by the method of analogues. Am J Epidemiol 158: 996–1006.
11. Rath T, Carreras M, Sebastiani P (2003) Automated detection of influenza epidemics with hidden markov models. In: Berthold M, Lenz H, Bradley E, Kruse R, Borgelt C, eds. Advances in Intelligent Data Analysis V. Berlin-Heidelberg: Springer. pp 521–532.
12. Yang P, Duan W, Lv M, Shi W, Peng X, et al. (2009) Review of an influenza surveillance system, beijing, people's republic of china. Emerg Infect Dis 15: 1603–1608.
13. Fleming D, Zambon M, Bartelds A, de Jong J (1999) The duration and magnitude of influenza epidemics: A study of surveillance data from sentinel general practices in england, wales and the netherlands. Eur J Epidemiol 15: 467–473.
14. Quénel P, Dab W (1998) Influenza a and b epidemic criteria based on time-series analysis of health services surveillance data. Eur J Epidemiol 14: 275–285.
15. Cowling BJ, Wong IOL, Ho LM, Riley S, Leung GM (2006) Methods for monitoring influenza surveillance data. Int J Epidemiol 35: 1314–1321.
16. Jiang X, Wallstrom G, Cooper GF, Wagner MM (2009) Bayesian prediction of an epidemic curve. J Biomed Inform 42: 90–99.
17. Polgreen P, Chen Z, Segre A, Harris M, Pentella M, et al. (2009) Optimizing influenza sentinel surveillance at the state level. Am J Epidemiol 170.
18. Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinki M, et al. (2009) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.
19. Valdivia A, López-Alcalde J, Vicente M, Pichiule M, Ruiz M, et al. (2010) Monitoring influenza activity in europe with google flu trends: comparison with the findings of sentinel physician networks – results for 2009–10. Euro Surveill 15: 1–6.
20. Wilson N, Mason K, Tobias M, Peacey M, Huang Q, et al. (2009) Pandemic influenza a (h1n1) v in new zealand: the experience from april to august 2009. Euro Surveill 14: 19386.
21. Googleorg (2003) Explore flu trends - United States. http://www.google.org/flutrends/us/#US-TX. Accessed February 13th 2012.
22. Langmuir A, Housworth J (1969) A critical evaluation of influenza surveillance. Bull World Health Organ 41: 393–398.
23. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle J (2000) Clinical Signs and Symptoms Predicting Influenza Infection. Arch Intern Med 160: 3242–3247.
24. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312: 447–451.
25. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2008) Detecting influenza epidemics using search engine query data. Nature 457: 1012–1014.
26. Censusgov (2002) Census 2000 US Gazetteer Files. http://www2.census.gov/census_2000/datasets/ Accessed February 13th 2012.
27. Thompson WW, Shay DK, Weintraub E, Brammer L, Bridges CB, et al. (2004) Influenza-associated hospitalizations in the united states. JAMA 292: 1333–1340.
28. Das A, Kempe D (2008) Algorithms for subset selection in linear regression. In: Proceedings of the 40th annual ACM symposium on Theory of computing. New York: ACM. pp 45–54.
29. Nemhauser GL, Wolsey LA, Fisher ML (1978) An analysis of approximations for maximizing submodular set functions–i. Math Program 14: 265–294.
30. Church R, ReVelle C (1974) The maximal covering location problem. Papers in Regional Science 32: 101–118.
31. Jolliffe I (2002) Principal Component Analysis. New York: Springer series in statistics.