# Generation and Analysis of Large Synthetic Social Contact Networks.

**Conference Paper** *in* Proceedings - Winter Simulation Conference · December 2009

DOI: 10.1109/WSC.2009.5429425 · Source: DBLP



**8 authors**, including:

Chris Barrett
Virginia Polytechnic Institute and State University
**145** PUBLICATIONS **3,288** CITATIONS
SEE PROFILE

Richard Beckman
Virginia Polytechnic Institute and State University
**118** PUBLICATIONS **9,560** CITATIONS
SEE PROFILE

Sritesh Kumar
Nagarjuna College
**135** PUBLICATIONS **3,694** CITATIONS
SEE PROFILE

Madhav Marathe
Virginia Polytechnic Institute and State University
**393** PUBLICATIONS **8,999** CITATIONS
SEE PROFILE

*Proceedings of the 2009 Winter Simulation Conference*
*M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.*

# GENERATION AND ANALYSIS OF LARGE SYNTHETIC SOCIAL CONTACT NETWORKS

Christopher L. Barrett
Richard J. Beckman
Maleq Khan
V.S. Anil Kumar
Madhav V. Marathe
Paula E. Stretz
Tridib Dutta
Bryan Lewis

Network Dynamics & Simulation Science Laboratory, Virginia Tech
1880 Pratt Drive, Building XV
Blacksburg, VA 24060, USA

## ABSTRACT

We describe "first principles" based methods for developing synthetic urban and national scale social contact networks. Unlike simple random graph techniques, these methods use real world data sources and combine them with behavioral and social theories to synthesize networks. We develop a synthetic population for the United States modeling every individual in the population including household structure, demographics and a 24-hour activity sequence. The process involves collecting and manipulating public and proprietary data sets integrated into a common architecture for data exchange and then using these data sets to generate new relations. A social contact network is derived from the synthetic population based on physical co-location of interacting persons. We use graph measures to compare and contrast the structural characteristics of the social networks that span different urban regions. We then simulate diffusion processes on these networks and analyze similarities and differences in the structure of the networks.

## 1 INTRODUCTION

The explosion in urban population in recent decades has resulted in very high social connectivity with a "small-world" structure. This has provided a perfect conduit for the spread of diseases, and as illustrated in the SARS epidemic from a few years back, diseases can spread on a global scale very quickly, and need quick response and interventions to prevent them from turning into large epidemics. Understanding the urban social-contact structure is critical for social scientists, urban planners, infrastructure companies, and governments, because a number of recent studies have shown that the spatial distribution of population in a city and mobility patterns of people have a significant impact on the disease transmission (Eubank et al. 2004). Additionally, social networks do not form and operate in isolation, and there is a significant amount of interaction and co-evolution between social and infrastructure networks (e.g., the transport and communication networks) (Stokman and Dorien 1997, Snijders et al. 2006, Steglich et al. 2007). Therefore, realistic models for social networks need to take other networks into account. There are very few tools or realistic social network models available for policy planners to understand the structure of such social networks, primarily because of the immense difficulty in collecting reliable data for social contacts. Some of the existing models involve small data sets, with a few thousands individuals (Newman 2003, Framingham Heart Study <http://www.framinghamheartstudy.org/index.html>, AddHealth <http://www.cpc.unc.edu/projects/addhealth>).

For many network applications, such as the Internet, the web graph and the power grid, researchers have realized the importance of detailed network modeling. Here again, the real network structure is not easily available, partly because of com-

mercial and security concerns, and a number of sophisticated methods have been developed to infer the network structure by indirect measurements. However, no such methods are known in the case of social contact graphs, because of the kinds of different information sources needed for building them. The goal of this paper is to develop a methodology to construct realistic social networks, using a combination of public and private data sets and statistical and large-scale agent based techniques. We have used these methods to construct synthetic population models for a number of urban regions in the United States. These population models have been the basis for a number of studies on public health and transportation policy planning, conducted for government agencies, e.g. (Eubank et al. 2005, Halloran et al. 2008). The detail and theory used in the construction of the above social contact networks is essential. In contrast to several recent results (Newman 2003), we show that activity based social contact networks generated as above differ from classical models of random networks – they are not *scale-free* or *small-world*, and have more complex degree distributions, but have higher local clustering than random graphs with the same degree distribution. Hence, decisions based on relative simplistic networks such as random networks might not be very accurate.

### 1.1 Related Work

Building realistic social network models has been of interest for researchers in many different areas for a long time. Many of the constructed networks have been restricted to specific activities, locations or demographics. Two data sets which involve extensive real data measurements collected over a number of years are (National Longitudinal Study on Adolescent Health), which collected extensive longitudinal data on school children over several years, and the Framingham Heart Study; both these data sets also provide demographic information for the individuals involved in the studies. Most other data sets available in literature only have information about the contact structure (i.e., no demographic information), e.g., the Enron email network (Enron Email Dataset), Karate club (Newman 2003), etc. Also, these networks usually capture contacts in the form of communications (e.g., emails), which are somewhat easier to observe or model, instead of physical contacts. Note that actual physical contacts are needed in order to study disease spread in these systems. There has been significant amount of research on inferring the structure of infrastructure networks, e.g., the Internet (Li et al. 2004, Seshadri et al. 2008), the Web graph (Kumar et al. 2000), recommender system graphs (Leskovec et al. 2006) and the power grid. Most of these networks share characteristics such as scale free degree distribution, high clustering, small diameter (Newman 2003, Barabasi and Albert 1999). Most of the social networks studied have very high robustness to both random and targeted attacks, while the infrastructure networks, such as the Internet have been found to be robust to random attacks, but highly vulnerable to targeted attacks (Eubank et al. 2004, Newman 2003, Barrett et al. 2007). The classical graph models, such as Erdos-Renyi do not exhibit such properties, and a number of more sophisticated random graph models have been developed (Newman 2003, Barrett et al. 2007). However, it has been observed that such random graph models can capture properties of realistic networks only to certain extent, and there has been a lot of interest in developing "first principles" methods for the Internet graph, which explicitly take into account the technological and economic issues arising in the construction of these networks (Li et al. 2004).

However, none of the network models in literature deal with large heterogeneous urban populations, which would be needed for understanding human disease spread. Collecting all the data needed to build such models is very difficult, because of privacy and security concerns. Therefore, any detailed population model has to be built by combining a variety of data sets. The population model we describe in this paper is the most refined model known, to our knowledge.

## 2 SOCIAL NETWORK CONSTRUCTION

We describe "first principles" based methods for developing synthetic urban and national scale social contact networks. Unlike simple random graph techniques, which attempt to match certain aggregate properties (e.g., degree and clustering coefficient distributions), these methods use over a dozen real world data sources and procedural knowledge about urban mobility and combine them with behavioral and social theories to synthesize networks, e.g., (Macy et al. 2002). We develop a synthetic population for the United States that models every individual in the population, though our techniques could be used in any region, with appropriate data and expert knowledge, thus making this a "first-principles" approach, analogous to the research done for in the case of the Internet (Li et al. 2004). Household structure and demographics are derived from U.S. Census data. Each synthetic individual is assigned a 24-hour activity sequence including geo-locations for each activity. A social contact network is constructed based on physical co-location of the interacting persons. We use graph measures to compare and contrast the structural characteristics of the social networks that span different urban regions. Our results show that realistic social contact networks: (*i*) are structurally different than synthetic networks generated using simple random processes, (*ii*) show interesting commonalities as well as differences as a function of the underlying urban region.

Our work builds on our earlier work in synthesis and analysis of large relational networks. Initial work was done under the TRANSIMS and NISAC projects (Barrett et al. 2001, Beckman et al. 1996, Eubank et al. 2004, Eubank et al. 2005, Bar-

rett et al. 2007) and more recently new methods have been developed under the Simfrastructure project. Our approach for synthesizing urban scale social contact networks involves the following steps (see (Eubank et al. 2005, Barrett et al. 2007) for more details). Step 1 creates a synthetic urban population by integrating a variety of databases from commercial and public sources into a common architecture for data exchange. The process preserves the confidentiality of the original data sets, yet produces realistic attributes and demographics for the synthetic individuals. The synthetic population is a set of geographically located people and households (referred to as a *proto-population*), each associated with demographic variables drawn from any of the demographics available in the census. Each synthetic individual is placed in a household with other synthetic people and each household is located geographically in such a way that a census of our synthetic population yields results that are statistically indistinguishable from the original census data, if they are both aggregated to the block group level. In Step 2, a set of activity templates for households are determined, based on several thousand responses to an activity or time-use survey. These activity templates include the sort of activities each household member performs and the time of day they are performed. Thus for a city - demographic information for each person and location, and a minute-by-minute schedule of each person's activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques; this information can be captured by a *dynamic social contact network*. See (Chowell et al. 2003) for analyses done using synthetic networks that were generated using our overall methodology for the city of Portland. Note that it is *impossible* to build such a network by simply collecting field data; the use of generative models to build such networks is a unique feature of this work. Recently Anderson et al. have persuasively argued the value of such an approach and proposed a similar method for constructing IP and ISP networks (Li et al. 2004, Seshadri et al. 2008).

A substantial effort has been spent on calibration and validation of our relational networks; see (Barrett et al. 2001, Chowell et al. 2003, Eubank et al. 2004, Barrett et al. 2007) for details. First, the design of the system is based on a formal theory of simulation called Sequential Dynamical Systems (Eubank et al. 2005, Barrett et al. 2003, Barrett et al. 2007). Various microscopic and macroscopic quantities produced by TRANSIMS have been validated in the city of Portland, including (*i*) traffic invariants such as flow density patterns and jam wave propagation; (*ii*) macroscopic quantities, such as activities and population densities in the entire city, number of people occupying various locations in a time varying fashion, time varying traffic density split by trip purpose and various modal choices over highways and other major roads, turn counts, number of trips going between zones in a city, etc. Results on population mobility and social network construction were presented and reviewed annually (Barrett et al. 2001). The results were also reviewed in the context of epidemic modeling as a part of a letter report by the National Academies and published in (Halloran et al. 2008).

The current paper extends our earlier work in several directions. First new sources of data and surveys are used for inferring these social networks. Second, using these methods, we have created synthetic social contact networks for the entire US and report in this paper our analysis of social networks that span urban regions of the country. The choice of these regions was based on covering various parts of US based on demographics, geography and city structure. Third, most of our early work was based on handling flat files. These solutions although adequate for constructing a single network were found to be inadequate in terms of desired diversity. We have thus developed efficient methods using current database technology. We also undertake a comparative analysis of the social networks that span urban and rural regions in the US. No such study has been reported prior to this work.

### 2.1 Synthetic Population Generation

The description in this section is based on (Barrett et al. 2005). Recall the definition of proto-populations described above; they can be generalized to represent a person, a vehicle, or an infrastructure element such as a hospital or a switch (Barrett et al. 2007). Here, we concentrate on creation of synthetic urban populations. Figure 1 shows a schematic diagram.

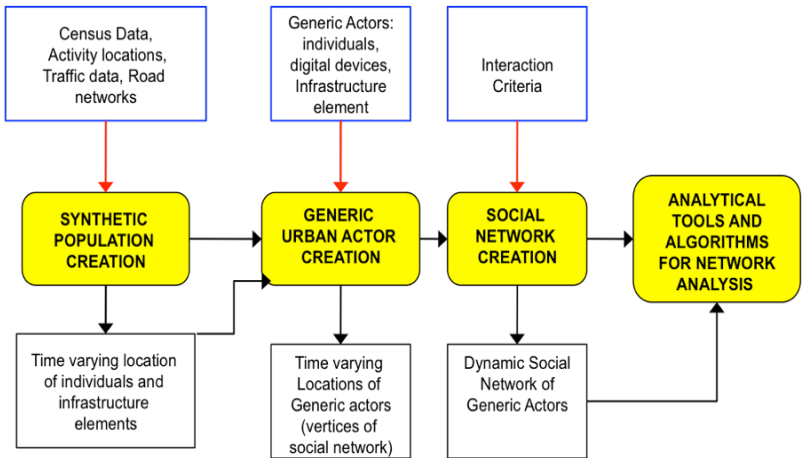


Figure 1: Synthetic social network generation. Generic actors are individuals or active devices that interact with each other.

Joint demographic distributions can be reconstructed from marginal distributions available in typical census data using an iterative proportional fitting (IPF) technique (Beckman, Baggerly, and McKay 1996). The synthetic population is statistically indistinguishable from the census data. Since they are synthetic, the privacy of individuals within the population is protected. The synthetic individuals carry with them a complete range of demographic attributes collected from the census data, including variables such as income level and age.

### 2.2 Social Network Construction

A set of activity templates for households is determined, based on several thousand responses to an activity or time-use survey. These activity templates include the sort of activities each household member performs and the time of day they are performed. Each synthetic household is then matched with one of the survey households, using a decision tree based on demographics such as the number of workers in the household, number of children of various ages, etc. The synthetic household is assigned the activity template of its matching survey household. For each household and each activity performed by this household, a preliminary assignment of a location is made based on observed land-use patterns, tax data, etc. This assignment must be calibrated against observed travel-time distributions. However, the travel-times corresponding to any particular assignment of activities to locations cannot be determined analytically. Using techniques in combinatorial optimization, machine learning and agent based modeling the populations, their activity locations, and their itineraries may be refined so as to be structurally and statistically consistent (Barrett et al. 2001).

The activities are modeled to take place at geographically located sites using data from commercially available databases of possible activity locations. Work, retail and recreation activity locations are derived from data from Dun & Bradstreet. School and college locations were constructed from data from National Center for Educational Statistics. Location choice for the activities is calibrated to the travel time data in the National Household Travel Survey (NHTS) (National Household Travel Survey 2001). NHTS contains data on the length and time of each trip taken by each individual in the survey. The purpose of the trips is also recorded. The trip purposes are denoted by H (home), W (work), S (shop), O (other) and C (school or college). Of these Home, Work, and School or College are considered "anchor" activities. These symbols are combined to designate a trip, for example HW is a trip from home to work.

The following methodology is used to assign the activity locations. The locations of all anchors in a home-to-home tour are determined first. The method for determining the location of the work or other "anchor" activity is relative to the home location or the last located "anchor" activity (Barrett et al. 2001).

For a home located at location $i$, the location of W for the trip HW is chosen from all possible locations $j$ with probability proportional to

$$P\{j|i\} = A_j e^{b_w D_{ij}}$$

where $A_j$ is an constant representing the "attractiveness" of location $j$ for work, $b_w$ is a calibration constant to be determined, and $D_{ij}$ is the distance from the home to the work location.

A similar relative probability equation holds for the location for non-anchor activities. Here the relative probability of choosing the location takes into account the location of the activity immediately preceding the activity and the location of the next anchor in the activity list.

A "gravity" methodology was used to determine the exact activity locations for each activity in individual person's activity pattern. In the gravity methodology, locations are chosen using a probability function that depends on the travel time or distance between the origin and the destination of the trip. The form of this probability function is

$$P\{D|O\} \propto e^{b \cdot Dist(O,D)},$$

where $D$ is a possible destination location, $O$ is the origin location, $Dist(O,D)$ is the distance or travel time between the two locations, and $b$ is a calibration constant that depends on the type of activity to be performed at location $D$. The values of $b$ were determined by statistically fitting the travel time and distance data from the NHTS to the gravity model.

### 2.2.1 Calibration Constants for Activity Location Process

Since we only have access to the Euclidian Distances between locations, the emphasis here is on the trip length distributions from the NHTS. Figure 2 shows the distribution of the length of trips for all of the trip types in NHTS excluding school/college trips. (Note: All of the trip length distributions given in this paper are derived by truncating the NHTS distances to consider only those trips less than 100 miles.) From this figure it is obvious that the HW and WH trips are longer than other trips involving non-work activities. Therefore, the location of the work activity is treated differently than that of the non-work activities. While somewhat different, the other trip length distributions are similar enough to consider them the same. Therefore, calibration constants, $b$, are developed for the location choice for work (see below as there are two such constants depending on the home location), and one constant is given for all other location choices.

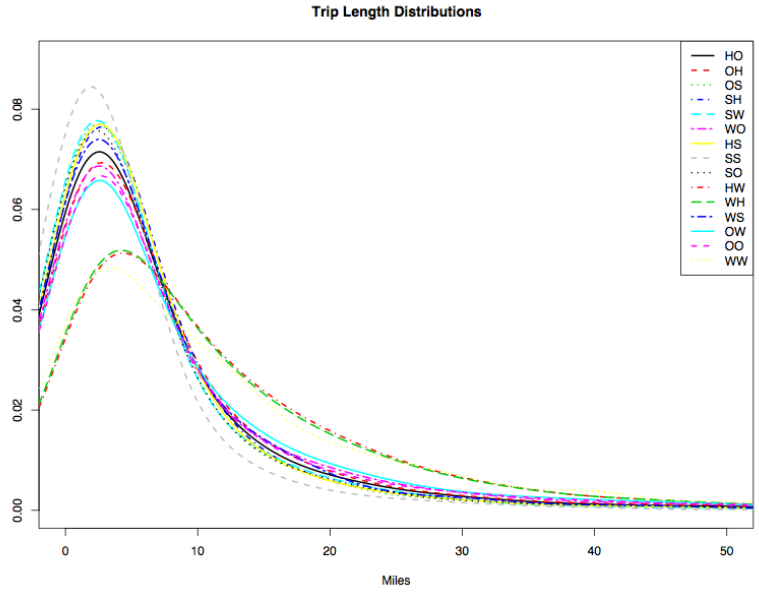


Figure 2: Trip length distribution by trip type

There is a common belief in the transportation community that the length of trips from home to work depends on the home location. If the home is in a suburban or rural area these trips tend to be longer. An analysis of the trip length distributions for the HW trips in the NHTS shows this to be true. Classification and Regression Trees (CART) were used to fit each of these trip distributions to a variety of independent variables that indicate the employment, population and household densities of the home locations, the region of the country of the survey household, an indicator of urbanization at the home location, and some household demographic data such as household income. The "best" CART fit split the HW travel distances into two distributions dependent on the value of the NHTS variable URBAN. In NHTS the variable URBAN has the following 4 values:

1. In an urban cluster

2. In an urban area
3. In an area surrounded by urban areas
4. Not in an urban area

The two distributions formed using this variable are the travel distances when URBAN=1 or 2 and URBAN=3 or 4. It is easy to see that this procedure splits the households into extremely urban and other areas.

A value for the URBAN variable needs to be assigned to every home location in the synthetic US population. This would be time consuming, so some of the other variables were considered for splitting the HW trips. One such is the NHTS variable HTHRESDEN, the number of housing units per square mile in the household's census tract. This variable was shown in the CART analysis to be almost as good as URBAN in splitting the trip distances from home to work.

The results from the CART fits show that there are two possible breakpoints for the variable HTHRESDEN. These are: 425 housing units/square mile, and 87.5 housing units/square mile. The trip length distributions resulting from these splits are shown in Figure 3.

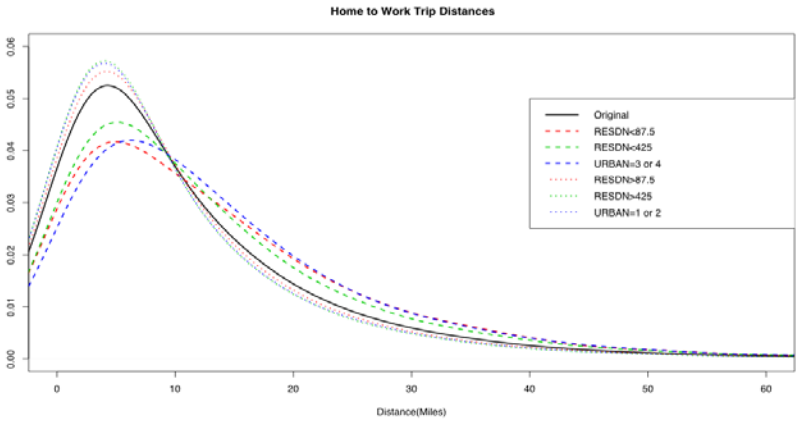


Figure 3: Trip length distribution for Home-to-Work trips based on housing units/sq mi in household census tract

The trip length distribution in black in the figure is the original distribution before splitting by the urbanization around the home location. The dotted and dashed lines are the two trip length distributions after the splits, where the blue lines are the split on the variable URBAN, the green show the split for HTHRESDN < or > 475 housing units/square mile, and the red show the split for HTHRESDN < or > 87.5 housing units/square mile. While any of these splits is adequate, the red splits, HTHRESDN < or > 87.5 housing units/square mile, seem to be the closest to the split using the "best" variable URBAN, so this split is used here.

To assign locations, both for work and the other activity types, initial values are determined for the calibration parameter $b$ in the equations $P\{j|i\} = A_j e^{b_w D_{ij}}$ and $P\{j|i\} = A_j e^{b_0 (D_{ij} + D_{jk})}$.

The following methodology is used for initialization of the parameter $b$. The average of the non-zero values of $A_j$ is computed and called $A$. At this time each of the attractor values, $A_j$, has the value 0 or 1, so the value of $A$ used here is 1. We use the Euclidian distance between two points as the value of $D_{ij}$ in the equation. The trip length data in the NHTS survey is the actual distance between two points, which by necessity is longer that the Euclidian distance. Therefore, the NHTS data is scaled to reflect this, and lacking any other information, this scaling is taken to be $\sqrt{2}$.

The median trip length, denoted by $\ell$, from the NHTS data is determined and the following equation is formed

$$0.5 = A e^{b\ell/\sqrt{2}}$$

Which leads to a solution for $b$:

$$b = \frac{\sqrt{2}}{\ell} \log \frac{0.5}{A}$$

The following assumptions are made in the solution for b above:

1. It is assumed the $A$ is always greater than or equal to 1.
2. The results given here are in miles. If other units are used (e.g. meters) the value of $b$ needs to be scaled to reflect this.

The median trip lengths for the home to work for the two cases, HTHRESDN < or > 87.5 housing units/square mile are:

1. For HTHRESDN < 87.5 housing units/square mile: Median = 5.0 miles, initial guess for $b$, $b$ = -.20
2. For HTHRESDN > 87.5 housing units/square mile: Median = 3.0 miles, initial guess for $b$, $b$ = -.33

The procedure for determining calibration constants for shop and other is the same as that for work with an additional factor of 2 to account for the sum of the two distances in the formula

$$P\{j|i\} = A_i e^{b_0(D_{ij}+D_{jk})}$$

and is used for the location of all non-anchor activities. The solution for $b$ for these non-anchor locations is

$$b = \frac{\sqrt{2}}{2\ell}\log\frac{0.5}{A}$$

In the first section it was determined that the trip length distribution of trips to and from non-anchor locations is almost independent of the type of trip being made. Therefore, the median used here to estimate the constant is the median trip length distribution for the home to shop combined with the home to other trips. The median trip length for these trips is 3.0 miles, so the estimate of $b$ is $b$ = -.16.

We consider 5 age groups for home to school trips and develop calibration constants for each age group.
As expected, elementary school children make shorter home to school trips than the others. We could fit one $b$ for elementary school location choice and one other one for all other school choices, but school is important for some of our models, so a separate value of $b$ is given for each of the age groups. As school is an anchor in our model, $b$ is estimated using the same procedure as used for work. The results are:

1. For 0 ≤ age ≤ 4: Median = 4 miles, $b$ = -.25.
2. For 5 ≤ age ≤ 12: Median = 2 miles, $b$ = -.49
3. For 13 ≤ age ≤ 15: Median = 3 miles, $b$ = -.33
4. For 16 ≤ age ≤ 18: Median = 4 miles, $b$ = -.25
5. For age > 18: Median = 3 miles, $b$ = -.33

### 2.2.2 The Interaction Network

Demographic information for each person and location, and a minute-by-minute schedule of each person's activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques. This forms the basis of the interaction network that can be abstractly represented by a (vertex and edge) labeled bipartite graph $G(P, L)$, where $P$ is the set of people and $L$ is the set of locations. If a person $p \in P$ visits a location $l \in L$, there is an edge $(p, l, \text{label}) \in E(G(P, L))$ between them, where label is a record of the type of activity of the visit and its start and end times. Each vertex (person and location) can also have labels (Figure 4). The person labels correspond to his/her demographic attributes such as age, income, etc. The labels attached to locations specify the location's attributes such as its $x$ and $y$ coordinates, the type of activity performed, maximum capacity, etc. Note that there can be multiple edges between a person and a location recording different visits.


Figure 4: Bi-partite graph of people and locations

### 2.3 Results

We used the methods above to select social networks from our United States synthetic population for urban regions in the United States. We discuss preliminary results for 3 networks here: Los Angeles, New York City, and Seattle. The social networks for these regions produce large graphs with millions of nodes and edges as shown in Table 1.

Table 1: Social network sizes for 9 urban regions

| Region | Number of Nodes | Number of Locations | Average Degree |
|---|---|---|---|
| Los Angeles | 16,228,759 | 3,201,621 | 56.596 |
| New York City | 17,876,290 | 4,348,939 | 53.713 |
| Seattle | 3,207,037 | 779,685 | 55.345 |

We treat the social networks as labeled graphs and measure characteristics of the graphs to discover similarities and differences in the graph structure of the regions. We discuss the following measures of these different graphs: Degree and Clustering Coefficient. Our main observations are the following:

1. There is remarkable similarity in the unlabeled measures over the different contact graphs. However, the disease dynamics on these graphs are very different, as discussed in the next section. Also, our work is the first exploration of the structure of labeled subgraphs in such contact networks.
2. These graphs differ significantly from other complex networks and random graph models studied in literature. For instance, the degree distributions satisfy a power law, as suggested in (Newman 2003, Barabasi and Albert 1999), in many complex networks. In fact, the degree distributions in our graphs do not satisfy power laws, and have multiple modes, which are closely related to sub-location sizes in our models.
3. The unlabeled graph measures do not seem to directly give insights into disease dynamics, and our results suggest that a closer study of demographic labels, as well as new structural measures are needed (such as the *vulnerability* measure) in order to understand disease dynamics. In contrast to literature in complex networks, which attempts to fully characterize the dynamics in terms of simple structural properties, we find that the vulnerability measure gives fundamentally new insights into the network structure.

Because of the scale of these contact graphs, even simple measures become challenging to compute (and most libraries of graph algorithms do not easily scale for such graphs). Also, the labeled structure leads to a rich new set of measures. Therefore, we need new sampling and streaming based methods for computing the properties of such large contact graphs.

#### 2.3.1 Graph Measures

Recall that $G=(P, L, E)$ denotes the labeled bipartite interaction graph that captures visits by people to different locations. This induces a person-person graph $G_P=(P, E_P)$ on the set of people, where there is a contact edge $(u,v) \in E_P$ if the individuals $u$ and $v$ come into contact at some common location $l \in L$. While other projections of this graph are also interesting, our main focus is the spread of epidemics, for which the projection $G_P$ seems most suitable. Let $N(v)$ denote the set of neighbors of node $v$, and let $\deg(v)=|N(v)|$ be the number of neighbors of $v$. The clustering coefficient of a node $v$ is defined as $|\{(w,w') \in E_P : w, w' \in N(v)\}| / (\deg(v)\cdot(\deg(v) - 1) / 2)$; thus, the clustering coefficient (CC) of node $v$ is the fraction of its neighbors that come into contact out of $(\deg(v)\cdot(\deg(v) - 1) / 2)$ possible pairs of neighbors. The $R_0$ (reproductive number) of a node $v$ is defined as the expected number of infections caused by node $v$ in one unit of time.

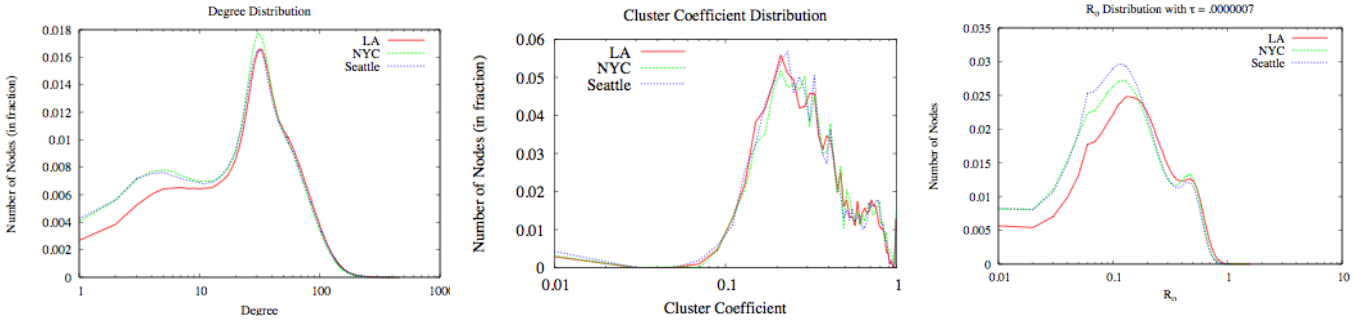


Figure 5: Degree, clustering coefficient, and Ro distributions

Figure 5 plots the frequency distribution of the $R_0$ values of all nodes; note that these are not very different. For homogenous, or complete mixing models, the $R_0$ distribution has significant impact on the disease dynamics. In the Figure 5, we plot the frequency distributions of all these quantities.

We count the occurrences of various non-overlapping subgraphs such as a clique, cycle, chain, or star (also called a template graphs) in a social contact graph. Two occurrences of a subgraph are non-overlapping if they do not share any edge or vertex of the graph. We observe that the counts of cliques and cycles in a social contact graph differ significantly from that in a random graph even when the degree distribution of the random graph is exactly as same as that of the social contact graph. A random graph with the same degree distribution is generated by shuffling the edges of the original graph as follows: randomly pick two edges of the graph and switch their end-points; repeat this process until all of the edges are shuffled. Figure 6 shows the counts of various template subgraphs in social contact graph and in a random graph with the same degree distribution. One significant difference is that the numbers of cliques of any size and smaller cycles in the random graph are almost zero. These graph measures demonstrate structural properties of our social contact networks. Understanding structure of a social contact network is important in understanding the disease dynamics in it.

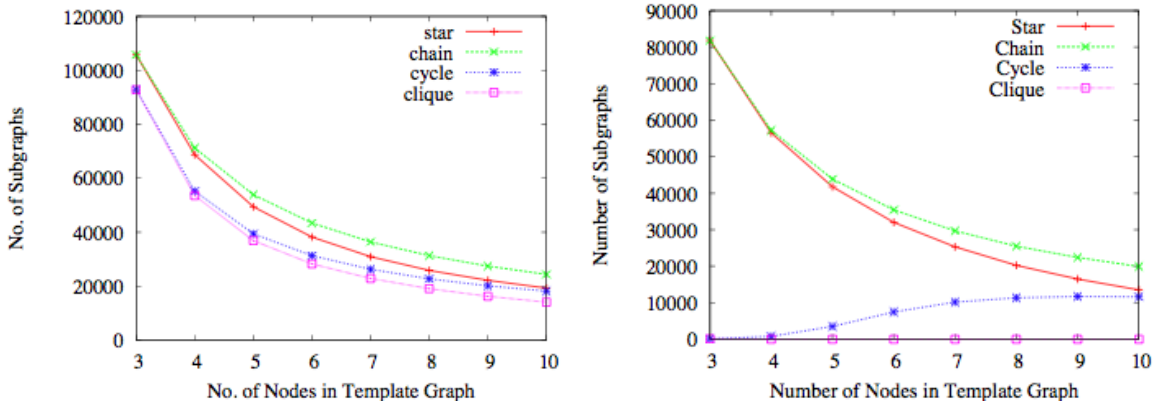


Figure 6: The number of occurrences of various subgraphs (star, chain, cycle, and clique) with the number nodes varying from 3 to 10 in a contact network of one of the cities and in a random network with the same degree distribution. Original contact graph (left), Random graph after shuffling the edges (right).

### 2.3.2 Diffusion Process on Social Networks

We simulate a diffusion process on the social contact networks of the regions; in this case influenza, that is transmitted across the graph (Bisset et al. 2009). Graph labels are the type of contact (work, home, school, shop, other) and the length of the contact. Well-known epidemiology measures of the spread of the disease such as the effective reproduction number (R0) are used to compare the social network graphs of each region. Although the social network for all regions was constructed using the same methods as described in the sections above, we find regional differences in the epidemiology measures.

The social contact networks can also be analyzed using any available demographic. Our analyses using age are presented in this paper. We consider 4 age groups:

- Preschool – age < 5
- School-age – age 5 – 18
- Adults – age 19 – 64
- Seniors – age 65+

Figure 7 shows the cumulative proportions of the populations infected with the influenza across the networks constructed for New York City, Los Angeles, and Seattle. The figure displays these attack rates for both the entire population and the populations broken into the four age groups given above. Two things are apparent from these figures. First, the proportion of the population infected in Los Angeles is greater than that of the other two regions. Second, a higher proportion of school-aged children, when compared to any other age group, are infected in all three regions.

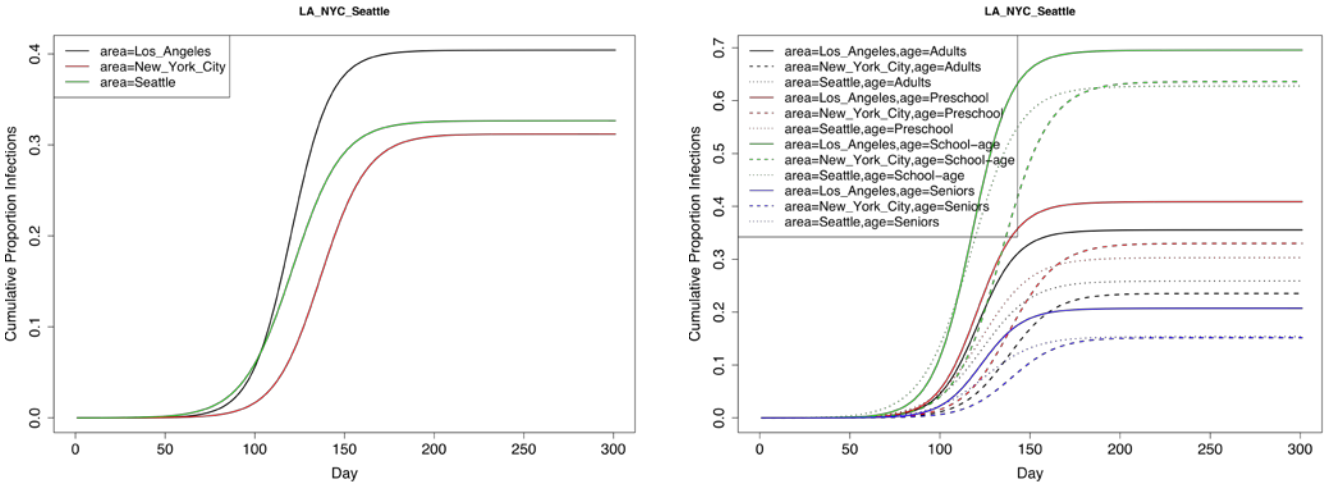


Figure 7: Cumulative proportion infections all ages (left) and daily cumulative proportion infections by age group (right) for NYC, LA and Seattle networks

Figure 8 shows, in histogram form, the R0 distributions. The degree distributions are weighted by the edge weights and represent the expected number of secondary infections caused by each person (node) in the network. The R0 distribution for Los Angeles in this figure is shifted to the right and hence one would expect a higher proportion of Los Angeles to be infected. We are investigating a similar analysis for the sub-population effects.

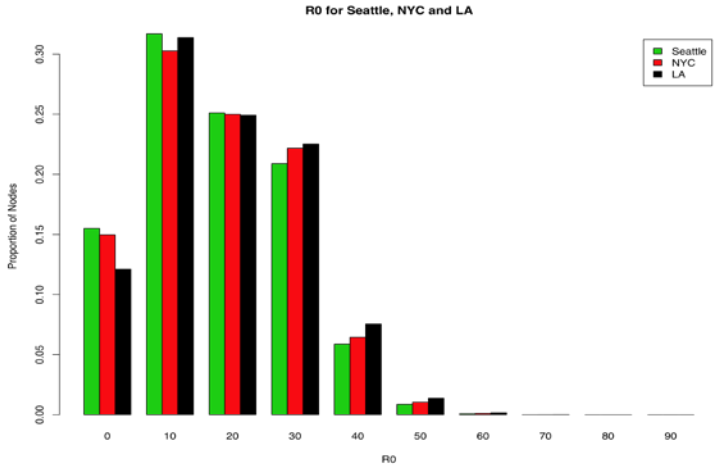


Figure 8: R0 distribution for NYC, LA and Seattle networks

## 2.4 Conclusions

We described a "first principles" approach for generating synthetic social contact networks spanning urban and rural regions in the US. Our methods are based on integrating various real world summary data sets and using appropriate social and behavioral theories to infer the relational networks. Database technology was used to automate and simplify many of the data fusion steps. New algorithms and their implementations were necessary to compute the structural properties of the ensuing social networks. These algorithms had to scale to be able to process networks with 5-15 Million nodes and 600 Million edges. These networks are labeled, dynamic and exhibit variability that reflects particular features of the city or the rural area. We show that the structure of these networks has significant impact on the dynamical processes on these networks; we use epidemics in urban regions to illustrate this. The work motivates a number of new research questions. These include, methods for inferring smaller sub-networks, more refined models for long distance travel and faster algorithms for measuring other structural attributes of large networks.

**ACKNOWLEDGMENTS**

We thank the members of the Network Dynamics and Simulation Science Laboratory (NDSSL) and our colleagues at Los Alamos National Laboratory for their collaborations, useful discussions and comments. The work reported here is based on joint effort by current and past NDSSL team. This work is partially supported by NSF Nets Award CNS-0626964, NSF Award CNS-0845700, NSF HSD Award SES-0729441, CDC Center of Excellence in Public Health Informatics Award 2506055-01, NIH-NIGMS MIDAS project 5 U01 GM070694-05, DTRA CNIMS Grant HDTRA1-07-C-0113 and NSF NETS CNS-0831633.

**REFERENCES**

Barabasi, A. and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286(5439):509-512.

Barrett, C. L., R. J. Beckman, K. P. Berkbigler, K. R. Bisset, B. W. Bush, K. Campbell, S. Eubank, K. M. Henson, J. M. Hurford, D. A. Kubicek, M. V. Marathe, P. R. Romero, J. P. Smith, L. L. Smith, P. L. Speckman, P. E. Stretz, G. L. Thayer, E. V. Eeckhout, and M.D. Williams. 2001. TRANSIMS: Transportation Analysis Simulation System. Unclassified Technical Report, No. LA-UR-00-1725, Los Alamos National Laboratory.

Barrett, C., S. Eubuank, and M. Marathe. 2006. Modeling and Simulation of Large Biological, Information and Socio-Technical Systems: An Interaction Based Approach. In *Interactive Computation: The New Paradigm,* D. Goldin, S. Smolka, and P. Wegner Eds. Springer Verlag.

Barrett, C., H.B. Hunt III, M.V. Marathe, S.S. Ravi, D.J. Rosenkrantz, and R.E. Stearns. 2003. Reachability problems for sequential dynamical systems with threshold functions. *Theoretical Computer Science* 295(1-3):41-64.

Barrett, C., K. Bisset, S. Eubank, V.S. Anil Kumar, M.V. Marathe, and H. Mortveit. 2007. Modeling and simulation of large biological, information and socio-technical systems: An interaction-based approach. In *Proceedings of the Short Course on Modeling and Simulation of Biological Networks*.

Beckman, R., K. Baggerly and M. McKay. 1996. Creating base-line populations. *Transportation Research Part A: Policy and Practice* 30(6):415-429.

Bisset, K., J. Chen, X. Feng, V.S. A. Kumar, and M. Marathe. 2009. EpiFast: A fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *Proceedings of the 23rd Annual International Conference on Supercomputing (ICS)*.

Chowell, G., J.M. Hyman, S. Eubank and C. Castillo-Chavez. 2003. Scaling laws for the movement of people between locations in a large city. *Physical Review E* 68:066102.

Eubank, S., H. Guclu, V. S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modeling disease outbreaks in realistic urban social networks. *Nature* 429(6998):180-184.

Eubank, S., V. S. Anil Kumar, M. Marathe, A. Srinivasan, and N. Wang. 2006. Structure of social networks and their impact on epidemics. In *Discrete Methods in Epidemiology*, Abello J, Cormode, G (ed.). *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 70:179-185.

Enron Email Dataset, <http://www.cs.cmu.edu/~enron/>. [Accessed April 9, 2009].

Framingham Heart Study. Available via <http://www.framinghamheartstudy.org/index.html>. [Accessed April 9, 2009]

Halloran, M.E., N. M. Ferguson, S. Eubank, I. M. Longini, D.A.T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, D. Wagener, R. Beckman, K. Kadau, C. Barrett, C. A. Macken, D. S. Burke, , and P. Cooley. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. In *Proceedings of the National Academy of Scences (PNAS)* 105(12): 4639-4644.

Kumar R., P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. 2000. The Web as a graph. In *Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*.

Leskovec, J., A. Singh, and J. Kleinberg. 2006. Patterns of Influence in a Recommendation Network. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

Li, L., D. Alderson, W. Willinger, and J. Doyle. 2004. A first-principles approach to understanding the internet's router-level topology. In *Proceedings of the 2004 ACM conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, New York, NY.

Macy, M. and R. Willer. 2002. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology* 28:143-166.

National Household Travel Survey. 2001. Available via <http://nhts.ornl.gov/download.shtml> [accessed April 8, 2009].

National Longitudinal Study of Adolescent Health (Add Health). Available via <www.cpc.unc.edu/projects/addhealth>. [Accessed April 9, 2009].

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45(2):167-256.

Seshadri, M., S. Machiraju, A. Sridharan, J. Bolot, J. Leskovec, and C. Faloutsos. 2008. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Snijders, T., C. Steglich, M. Schweinberger, 2006. Modeling the co-evolution of Networks and Behavior. *Longitudinal models in the behavioral and related sciences,* edited by Kees van Montfort, Han Oud and Albert Satorra; Lawrence Erlbaum.

Steglich, C., T. Snijders, and M. Pearson. 2007. Dynamic Networks and Behavior: Separating Selection from Influence. Technical report available at <http://stat.gamma.rug.nl/snijders> [Accessed April 9, 2009] .

Stokman, F. and P. Dorien. 1997. *Evolution of Social Networks*, Amsterdam, Gordan and Breach.

## AUTHOR BIOGRAPHIES

**CHRISTOPHER L. BARRETT** is a Professor in the Dept. of Computer Science and the Director of Network Dynamics and Simulation Science Laboratory, in Virginia Bio-Informatics Institute at Virginia Tech. His research interests are in modeling and simulations, computational neuroscience, discrete dynamical systems, computational epidemiology and communication networks. His email is <cbarrett@vbi.vt.edu>.

**RICHARD J. BECKMAN** is a Senior Research Associate in the Network Dynamics and Simulation Science Laboratory in Virginia Bioinformatics Institute at Virginia Tech. He is a Fellow of the American Statistical Association. His research interests include statistical characterizations of complex systems and the study of sensitivity and uncertainty in the outputs from computer simulations. His email is <rbeckman@vbi.vt.edu>.

**MALEQ KHAN** is a Postdoctoral fellow in the Network Dynamics and Simulation Science Laboratory in Virginia Bio-Informatics Institute at Virginia Tech. His research interests are in network science, randomized and distributed algorithms, discrete mathematics, and graph mining. His email is <maleq@vbi.vt.edu>.

**V.S. ANIL KUMAR** is a Assistant Professor in the Dept. of Computer Science and a Senior research Associate in the Network Dynamics and Simulation Science Laboratory, in Virginia Bio-Informatics Institute at Virginia Tech. His research interests are in network science, communication networks, design and analysis of algorithms, discrete mathematics and computational epidemiology. His email is <akumar@vbi.vt.edu>.

**MADHAV V. MARATHE** is a Professor in the Dept. of Computer Science and the Deputy Director of Network Dynamics and Simulation Science Laboratory, in Virginia Bio-Informatics Institute at Virginia Tech. His research interests are in network science, computational science and engineering, communication networks, theoretical computer science and computational epidemiology. His email is <mmarathe@vbi.vt.edu>.

**PAULA E. STRETZ** is a Senior Research Associate in the Network Dynamics and Simulation Science Laboratory in Virginia Bio-informatics Institute at Virginia Tech. Her interests include techniques for construction of synthetic populations and computer simulations using synthetic networks. Her email is <pstretz@vbi.vt.edu>.

**TRIDIB DUTTA** is a PhD candidate in the Dept. of Computer Science and in the Network Dynamics and Simulation Science Laboratory, in Virginia Bio-Informatics Institute at Virginia Tech. His research interests are in algorithms, data and graph mining, algebra and combinatorics. His email is <tridib@vbi.vt.edu>.

**BRYAN S. LEWIS** is a PhD candidate in the Genetics, Bioinformatics, and Computational Biology Department at Virginia Tech. He received his MPH from the University of California, Berkeley. His research interests include infectious disease modeling, public health policy, statistical analysis, and graph theory. His email is <blewis@vbi.vt.edu>.