Comparison and Validation of Synthetic Social Contact Networks for Epidemic Modeling

(Extended Abstract)

Huadong Xia, Jiangzhuo Chen, Madhav V. Marathe, Samarth Swarup
Network Dynamics and Simulation Science Laboratory,
Virginia Bioinformatics Institute,
Virginia Tech,
Blacksburg, VA, USA.
{xhd, chenj, mmarathe, swarup}@vbi.vt.edu

ABSTRACT

We describe the synthesis of detailed social contact networks of Delhi, India, and Los Angeles, USA, for urban-scale epidemiological simulations. The network synthesis is done by combining information from multiple data sources, since social contact information cannot be obtained through direct surveys. We compare the two networks on various structural and dynamical metrics. Through the comparison between the two cities, we show important similarities and differences between urban regions in different parts of the world.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation

Keywords

Computational Epidemiology, Synthetic Populations, Social Contact Networks

1. INTRODUCTION

This paper describes the construction of large-scale, data-driven synthetic social contact networks for epidemic simulations. In a social contact network, the nodes represent people, and edges represent interactions between people who come into close enough contact with each other to transmit an infection. This is the network over which an epidemic propagates, it cannot be obtained through direct surveys, because most people do not know all the people they come into contact with during a day.

Our approach is to construct an approximation to the social contact network by combining data about demographics, activity patterns, and activity locations. We describe the construction of synthetic social contact networks for the cities of Los Angeles, USA, and Delhi, India. These are large populations, and their social contact networks consist of millions of nodes and hundreds of millions of edges. Furthermore the networks are quite irregular and are not well-described by simple network models.

A crucial question is how to assess the quality of the constructed networks. This is broadly a question of validation. However, validation of such complex models is not simply a matter of comparing epicurves against real data, since it is very easy to match a sequence of numbers of infections by tweaking any of a large number of parameters.

Our approach here is to do a detailed comparison of structural and dynamical metrics on the two networks to determine the differences between them. This serves as a kind of cross-validation, since the two networks are generated from different data sources. Additionally, the process for generating the Los Angeles network follows relatively mature technology using high-quality data sources, and has been used in multiple prior studies that have gone through rigorous peer review [1,3]. By comparing the Delhi network with the Los Angeles network, we can build trust in the Delhi model if the differences between it and the Los Angeles network are explainable in terms of cultural and demographic differences between the two cities.

2. METHOD

We propose two methods to generate synthetic populations and networks for Delhi and Los Angeles. Both methods consist of the following broad steps: (i) synthesize a baseline population with a detailed individual structure and the same aggregate statistical properties of the real population; (ii) assign each individual a reasonable activity schedule based on mobility survey; (iii) create locations in the region where synthetic people can take their activities.

The two methods differ in specific models used in each step because the data sources of the two cities have different format or quality. First, subjective surveys regarding individual behavior in residential areas are used to synthesize the Delhi network but not for the Los Angeles network. This is because about 40% of the population in Delhi do not travel on a daily basis and have activities mainly in their home and neighborhood. Second, some data sets used to synthesize the Los Angeles network are more refined than the ones used to construct the Delhi network, for example the activity survey data. In the Delhi network, no activity survey is available for the Delhi population. Therefore, a travel survey for another Indian city, Thane, is utilized. The

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

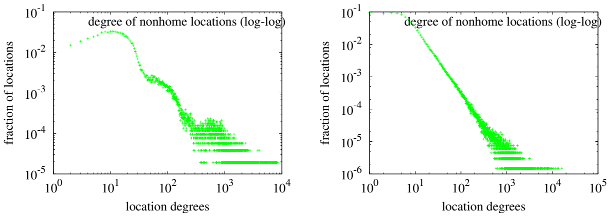
two networks constructed, therefore, reflect differences not only for the populations themselves but also those residing in the data sources.

Validation: We have made extensive efforts to validate our models; see [2] for an in-depth discussion. This includes: (i) data validation: matching diverse measured data sources, such as properties reported in the census, traffic, and people's activities, (ii) functional validation: ensuring the synthetic network is based on accepted social theories and data integration techniques, and (iii) structural validation: we use approaches from statistical physics and complex systems to show several emergent properties are consistent with observed system-level phenomena.

3. COMPARISON OF THE TWO NETWORKS

We compare the networks at various levels. Different from traditional approaches like ERGM models, realistic social contact networks are irregular, unstructured and dynamically changing, therefore difficult to measure or describe with any single metric. We describe a number of metrics for comparing the two networks. These metrics are divided into four classes: (i) metrics that capture the features of the population, built infrastructure and their layout (network labeling structure); (ii) network level metrics that capture the structural features of the dynamic social contact network; (iii) dynamical features that capture the epidemic dynamics over the networks; and (iv) policy metrics that capture the effect of controls. We briefly show some results below.

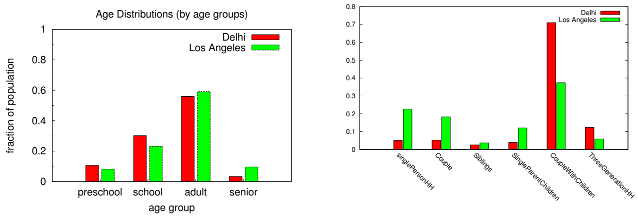
Person-location networks G_{PL}. The degree distribution of the people-location graph is plotted in Figure 1, wherein the two networks differ in details but both reveal a power law like degree distribution. This large-scale structure is documented frequently in the literature [4] (thus a validation of our networks).



(a) the Delhi network (b) The Los Angeles network

Figure 1: Degree distribution for people-location networks

Comparison of Epidemic Dynamics. Epidemic simulation shows that the internal interactions between subpopulations are very different in the two networks (Figure 3). For example, in the Los Angeles network, preschool children are a little bit more vulnerable than average (the red line), but in the Delhi network, they are the most resistant subpopulation. The difference comes from the different structural role the subpopulation plays in the two networks. Most preschool children in the Delhi network stay at home and thus correspond to high clustering low degree nodes, contributing to their low vulnerability. Children in the Los Angeles network go to daycare together thus representing a very different mixing pattern. The different mixing, together with different demographic structure and household structure in the two cities (Figure 2) shape the epidemics as shown. In addition, the school children are the most vulnerable subpopulation in both networks. These structural features may help us in designing effective intervention strategies.



(a) Subpopulations by age-group.

(b) Household structure comparison.

Figure 2: A comparison of the synthetic populations of Delhi and Los Angeles on some demographic measures.

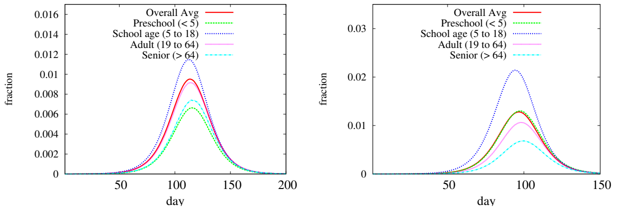


Figure 3: Epidemic curves show subpopulation epidemics in the Delhi network (left) and the Los Angeles network (right) when $R_0 = 1.35$. Both populations are partitioned to four groups based on age: preschool, school age, adult, and senior. Each dashed curve shows the fraction of people in that subpopulation infected on each day when there is no intervention.

4. CONCLUSION

We have introduced a number of metrics to describe and compare large social contact networks. Some metrics are well studied in the literature while others are new and capture the specific aspects of the networks. Our analysis reveals interesting differences and similarities between the networks.

5. REFERENCES

[1] C. Barrett, R. Beckman, M. Khan, V. Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *Winter Simulation Conference*, pages 1003–1014, 2009.

[2] R. Beckman, K. Channakeshava, F. Huang, J. Kim, A. Marathe, M. Marathe, G. Pei, S. Saha, and A. K. S. Vullikanti. Integrated multi-network modeling environment for spectrum management. *Selected Areas in Communications, IEEE Journal on*, 31(6):1158–1168, 2013.

[3] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[4] S. Eubank, V. A. Kumar, M. Marathe, A. Srinivasan, and N. Wang. Structure of social contact networks and their impact on epidemics. *AMS-DIMACS Special Issue on Epidemiology*, pages 181–213, 2006.