Towards "Live" Synthetic Populations for Large-scale Realistic Multiagent Simulations

(Doctoral Consortium)

Nidhi Parikh
Network Dynamics and Simulation Science Lab,
Virginia Bioinformatics Institute,
Virginia Tech,
Blacksburg, VA, USA
nidhip@vbi.vt.edu

ABSTRACT

Synthetic populations attempt to capture population dynamics of a geographic region and hence are widely used in large-scale multiagent applications simulating real-world phenomena. However, current synthetic populations are mostly static — individuals are assumed to perform same daily routine every day.

My thesis aims at taking the first step towards making it a "live" synthetic population that would update automatically to reflect changes in the real population, by incorporating information from social media and other online data resources. As an initial step, I have extended synthetic population model for Washington DC metro area to include transient (tourists and business travelers) population. This is done by combining data from various online and offline data resources by hand. This subpopulation which keeps changing with time, has also shown to have an important effect on disease dynamics of the city. Next, I propose to use information from social media to improve activity patterns of individuals using hidden semi-Markov model.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

Synthetic Population, Social Media, Hidden Semi-Markov Model

1. INTRODUCTION

Many multiagent applications require knowing geographic location, activity patterns, and demographics of people, for example, in epidemic simulations, a person could only be infected by other infected person present at the same location at the same time. Similarly, the ability of an individual to buy a vaccine depends upon his demographics factors like age and household income.

Synthetic populations represent individuals along with their demographic (i.e., age, income) and mobility related (i.e., type of activity, location) information. They capture population dynamics of a city which makes them suitable for large-scale multiagent simulations like disease outbreak [3], social contagion [1], and disaster recovery and response [6].

At the Network Dynamics and Simulation Science Laboratory (NDSSL), detailed, high fidelity synthetic populations [2] have been created for the United States by combining data from multiple resources like the American Community Survey, the National Household and Travel Survey, the National Center for Education Statistics, and geospatial data. However, this population is mostly static, i.e., it includes information about only residents and residents are assumed to perform the same set of activities every day.

In reality, dynamics of the city also depend upon visitors. Also, individuals usually follow different routines for weekdays and weekends, which could further be changed due to occasional change in plans, occurrence of events like game, festival in the city or due to some change in the environment. In an ideal scenario, we would like to have a "live" synthetic population which is updated on the fly as it collects information about the world which could probably be done by observing news, events or other online resources. However, this is a very broad goal and current synthetic populations are no where near to it. My thesis focuses on taking the first step towards this by incorporating information from social media and other online data sources. I have started by creating transient population, including tourists and business travelers for Washington DC metro area by combining various data sources. Epidemic simulations show that this changing subpopulation does play an import role in disease spread [5]. Recently, due to increased use of social media, a large amount of data is becoming available about people's activities and visits to various locations. So next, I propose to use Twitter data to create activity patterns for synthetic populations. The methods used to model transient population and proposed for using Twitter to identify activity patterns are described below.

2. MODELING TRANSIENT POPULATION

The methodology used to generate transient population [5] broadly follows the same for the current synthetic pop-

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

ulation. Destination DC[1] provides some demographic distributions for tourists and business travelers. Tourists individuals are grouped into parties of individuals traveling together. Within a party, these demographic distributions are not independent of each other, i.e., a married couple is more likely to travel together. Hence, we created a set of rules about the party structure and then use sampling without replacement in combination with these rules to generate synthetic transient individuals with demographics. Each party is then assigned a hotel location preferentially nearer to downtown. All individuals in a party are assumed to travel together and hence are assigned the same set of activities to perform during a day. Each activity is represented by the type of activity (i.e., hotel, tourism, work in case of business travelers), start time, duration, and location. Hotel and activity locations have been identified from D&B data using SIC (Standard Industrial Classification) codes. Activity location assignment is calibrated to match visit counts at Smithsonian Institution locations which are major tourist destinations in Washington DC.

We simulated a flu-like disease outbreak to evaluate the effect of transients on resident number of infections. It is assumed that every day approximately 20% of the transients leave the city and new transients with the exact same demographics replace them. The results show that there are more resident infections when transients are considered and disease also peaks about 7 days earlier.

3. IDENTIFYING ACTIVITY PATTERNS USING TWITTER

The activity templates that we currently use for the synthetic population are derived from the National Household Travel Survey data, and consist of following types of activities: home, work, school, shopping, other. Each synthetic individual is assigned a daily schedule of activities, which is written as a list of (activity type, start time, duration) tuples. Individuals are assigned activity templates based on their demographics.

To extend this to deriving activity schedules from social media data, the following conceptual model is used. During a day, an individual performs a sequence of activities. During these activities, he generates some output on social media, such as tweets on Twitter which provide information about the type of activity he is engaged in. Our goal is to infer his daily activity schedule from these tweets.

This problem could be formalized as a Hidden Semi-Markov Model (HSMM). An HSMM is an extension of a Hidden Markov Model (HMM) with the underlying process being semi-Markovian with variable duration for each state. It is defined by a state transition matrix, an observation distribution, and a duration distribution associated with each state. Each state can emit a series of observations depending upon the time spent in it. For our problem, the activity type corresponds to the hidden state, the duration of this hidden state corresponds to the time spent doing that activity and tweets correspond to the observations generated by the hidden state. An observation is represented by the features extracted from the tweet text as well as the geo-tags and time stamps.

In a general HSMM [8], there is an observation at each time step or whenever a state transition happens. However, here some observations could be missed, specifically there may be some activities for which there is no tweet (observation). But if it is assumed that an individual follows the same routine on weekdays or has fixed routines for each day of the week, then as data is collected over a long period of time (a few months), there are multiple observation sequences available (with some missing data). The task then is to combine these observation sequences using HSMM. There are some models proposed to deal with missing data and multiple observation sequences [7, 4] which will be evaluated and tweaked appropriately to fit this problem. I also plan investigate if prior knowledge about activity patterns from NHTS data could be used to reduce the complexity of learning the model either by providing constraints or by providing a starting point to guide the search for a good model.

Once HSMM model (initial state probabilities, transition probabilities, and duration distribution) is learned for a user, it could be used for generating activity template for a synthetic individuals with similar demographics.

[1] http://washington.org

4. REFERENCES

[1] A. Apolloni, K. Channakeshava, L. Durbeck, M. Khan, C. Kuhlman, B. Lewis, and S. Swarup. A study of information diffusion over a realistic social network model. In *The IEEE International Symposium on Social Computing Applications*, Vancouver, Canada, August 29-31 2009.

[2] C. Barrett, R. Beckman, K. Berkbigler, K. Bisset, B. Bush, K. Campbell, S. Eubank, K. Henson, J. Hurford, D. Kubicek, M. Marathe, P. Romero, J. Smith, L. Smith, P. Speckman, P. Stretz, G. Thayer, E. Eeckhout, and M. D. Williams. TRANSIMS: Transportation analysis and simulation system. Technical Report LA-UR-00-1725, Los Alamos National Laboratory, 2001.

[3] C. L. Barrett, K. R. Bisset, J. P. Leidig, A. Marathe, and M. V. Marathe. An integrated modeling environment to study the coevolution of networks, individual behavior and epidemics. *AI Magazine*, 31(1):75–87, 2010.

[4] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon. Information integration over time in unreliable and uncertain environments. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 789–798, New York, NY, USA, 2012. ACM.

[5] N. Parikh, S. Shirole, and S. Swarup. Modeling the effects of transient populations on epidemics. In *The AAAI Fall Symposium on Social Networks and Social Contagion*, Arlington, VA, Nov 2012.

[6] N. Parikh, S. Swarup, P. Stretz, C. Rivers, B. Lewis, M. Marathe, S. Eubank, C. Barrett, K. Lum, and Y. Chungbaek. Modeling human behavior in the aftermath of a hypothetical improvised nuclear detonation. In *Proc. AAMAS*, May 2013.

[7] S.-Z. Yu and H. Kobayashi. A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking. *Signal Processing*, 83(2):235–250, 2003.

[8] S. zheng Yu. Hidden semi-markov models. *Artificial Intelligence*, 2010.