Simulating Social Information Diffusion using a Synthetic Population

Andrea Apolloni, Karthik Channakeshava, Lisa Durbeck, Maleq Khan, Chris Kuhlman, Bryan Lewis, and Samarth Swarup

Network Dynamics and Simulation Science Lab, VBI, Virginia Tech, Blacksburg, VA 24061

Abstract

Sociological models of human behavior can explain population-level phenomena within social systems; computer modeling can simulate a wide variety of scenarios and allow one to pose and test hypotheses about the social system. In this work, we model and examine the spread of information through personal interaction in a simulated socio-technical network that provides a high degree of realism and a great deal of captured detail.

We use a probabilistic model to decide whether two people will converse about a particular topic based on their similarity and familiarity. Similarity is modeled by matching selected demographic characteristics, while familiarity is modeled by the amount of contact required to convey information. We report our findings on the effects of familiarity and similarity on the spread of information over the social network. We resolve the results by age group, daily activities, time, household income, household size and examine the relative effect of these factors.

For informal topics where little familiarity is required, shopping and recreational activities predominate; otherwise, home, work, and school predominate. We find that youths play a significant role in spreading information through a community rapidly, mainly through interactions in schools.

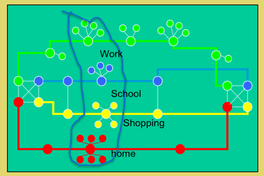
Generating a Synthetic Population

A synthetic population consisting of individuals and households is generated from census data. Locations are assigned to each household according to the census block information and are geolocated using the NavTeq street data. Demographics are assigned to each individual in the population based on household data.

Using data from activity surveys, a time ordered daily schedule of activities is matched with household demographics. Each activity sequence is then associated with a member of the household subject to the member's demographics. Each activity is denoted with a start time and duration, location where the activity is performed and an activity type (work, shopping, school etc.). Activity locations are derived from the survey based on distance traveled for particular activities. Each location, based on activity type is assigned within a certain distance from the household location. An individual's demographic and the attractors associated with locations are used to assign a location for the individual's activity.

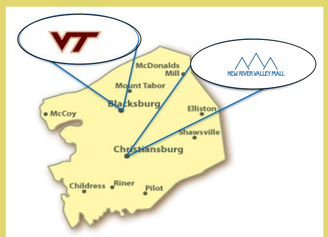
Generating a Social Interaction Network

The social network is created by using the interactions of the synthetic population at the activity locations. We use the activities to determine the occupancy of each location and model *sub-locations* to determine the interactions within each location. The number of sub-locations is determined by the occupancy of the location at every unit of time and we assume a certain occupancy for every sub-location (for example, 25 individuals per sub-location). The schedules specify the start and duration of contact for each link that is formed between individuals in each sub-location. We assume that all the individuals within the same sub-location form a clique. In addition, individuals have particular circadian periods, with different sleep times and duration, which depend on factors such as age and sex. Sleeping time and duration are considered as independent characteristics. Links are considered inactive if at least one of the two individuals is sleeping.

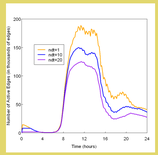


The New River Valley Region

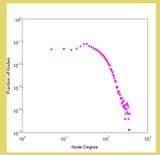
Population: 74,360
Links created: ~1,800,000



New River Valley is located in the state of Virginia, around the towns of Blacksburg and Christiansburg.



Number of active links over the course of a day. For definition of *ndt*, see next panel.



Degree distribution of the union graph.

Age	%	Household Income	%	Household Size	%
0 - 18	20	0 – 25k	33	1	11
18 - 35	39	25k – 50k	30	2 - 3	52
35 - 64	32	50k – 75k	20	≥ 4	37
> 64	9	> 75k	17		

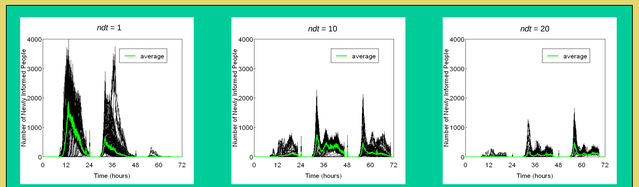
Major demographic strata in the population.

Diffusion of Information in the Synthetic Population

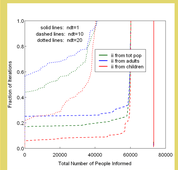
The type of information transmitted depends on the duration of contact.

Our information diffusion model makes the following assumptions:

• Individuals can interact only when they are collocated.



Epicurves



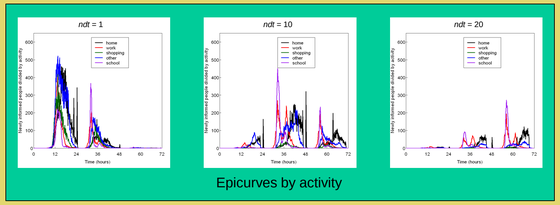
Final attack size

•The probability with which an individual conveys information to another individual is related to the similarity between the individuals, where similarity is based on demographic strata. The probability increases with increasing demographic overlap.

• Information transmission requires contact of a minimum duration, and the probability of transmission increases monotonically with the duration of contact beyond this minimal value. This simulates the fact that information transmission is not always immediate, but varies with the topic, i.e., some subjects might be broached only in a lengthy interactions, whereas others might be discussed even in short encounters. This is formalized as a *node duration threshold* (*ndt*), with a unit of 10 minutes, i.e., *ndt* = 1 means a pair of nodes that is in contact for less than 10 minutes has zero probability of transmitting information.

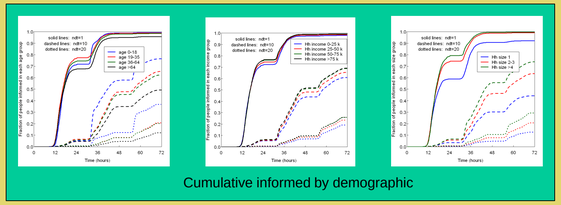
The figures above show the epicurves for information transmission over 3 days of simulated time, for three different values of *ndt,* and the final number of informed people (the final "attack size"). The epicurves are from 100 independent runs, along with their mean.

Diffusion by Activity and Demographic



Epicurves by activity

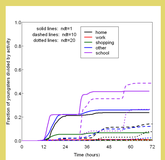
The figures above show the number of newly informed people with respect to the activity type. Increasing ndt changes the way people receive information due to the pruning of weak links. When ndt = 1, a majority of people are informed while recreating, shopping or at home. Increasing ndt causes information spread at school, home, and work to predominate. Further, increasing ndt shifts the peak of newly informed to day 2 and day 3.



Cumulative informed by demographic

The figures above show the fraction of people informed in each demographic stratum as a function of time. For ndt = 1 all the people present within the value range for each demographic factor are informed after ≈ 40 hours. Further, age and Hh income show a similar trend for all value ranges, whereas Hh size = 1 has a slower spread. Similarity across value ranges in middle plot indicates that Hh income does not play a differentiating role in the process.

Think of the Children



The figure on the left shows the fraction of informed youngsters at different activity locations. For youngsters, communication occurs mostly at school, independent of the ndt values. So, more youngsters get informed at school than at home. This is due to the fact that the probability of interaction with similar individuals is higher, and the duration of interaction is long enough to spread the information (strong ties).

Due to these long-duration interactions at school and home, youngsters form a strongly connected backbone of the social interaction network, making it hard to fragment the network by pruning short-duration interactions alone.

Conclusions

In this work, we have presented an interaction-based approach for studying the spread of rumors in a synthetic population under realistic conditions.
From our simulations, we conclude that if the information to be transmitted requires little familiarity between the individuals and the information can be transmitted in short conversation (say, with duration 10-20 minutes) recreation and shopping are activities where information spread is greatest. Otherwise, locations where individuals have higher familiarity and longer periods of time for conversation, such as home, work, and school, predominate. We also find that youngsters get informed mostly at school.

NICO Complexity Conference, Northwestern University, Chicago, September 2009