Generating social networks of intimate contacts for the study of public health intervention strategies

Courtney D. Corley, Lindsey Brown, Armin R. Mikler
University of North Texas
Department of Computer Science and Engineering
Denton, TX
{corley,lmb0105}@unt.edu,mikler@cse.unt.edu

Diane J. Cook
Washington State University
School of Electrical Engineering and Computer Science
Pullman, WA
cook@eecs.wsu.edu

Karan Singh
UNT Health Science Center
Department of Biostatistics
Ft. Worth, TX
ksingh@hsc.unt.edu

Abstract

Sexually transmitted diseases and infections are, by definition, transferred among intimate social settings. Although the circumstances under which these social settings are established and maintained may vary, the common prerequisite remains an intimate level of social atmosphere. For this reason, the development of sexually transmitted disease mathematical and computational models must utilize a precise and efficient social networking tool. This paper describes a computational simulator created to embody the intimate social networks related to the transmission of sexually transmitted diseases and infections for the utilization by public health professionals to facilitate evaluation of targeted intervention strategies.

1. Introduction

Sexually transmitted diseases and infections are a significant and increasing threat among both developed and developing countries around the world, causing varying degrees of mortality and morbidity in all populations [5]. The rates of prevalence of curable sexually transmitted diseases and infections are highest among the most developed countries, with a quarter of these conditions occurring within the 13-19 age range [6]. The responsibility of halting the dissemination of these conditions lies upon the shoulders of professionals within the public health industry. In order to properly and effectively use funding and resources, these individuals must have reliable tools to help predict the most appropriate means of intervention strategies.

In this paper we first describe the general algorithm used in our simulation for generating a contact driven network with a specific degree distribution, disease dynamics and for creating a dynamic population. We then describe in detail how our simulator connects the bipartite network with a predetermined degree distribution, minimizing the number of unresolved degrees. Next, several methodologies to analyze clustering in bipartite graphs are introduced. The networks generated by our simulator are then analyzed using the graph statistics introduced previously. Next, a sample case study is described that demonstrates our simulators capabilities. This paper concludes with describing future work to further develop our simulator.

Previous work employing social network schemes has varied in context. The EPISIMS computational analysis tool, created at the University of Maryland in conjunction with the Los Alamos National Laboratory, estimates social networking based on the transportation patterns evident within the target city, Portland, Oregon [7]. This computational model may be used to handle diverse social networking in regards to the transmission of infectious disease agents. Public health officials may utilize this model to help predict where preventive measures, including quarantine and vaccination, would be most useful and cost effective within their populations. Other avenues of previous social models of sexually transmitted diseases and infections have included the categorization of individuals into groups based on the differing stages of infection of each disease condition versus demographic factors such as age, sex, and geographic location [1, 16, 17].

Providing social networks for sexually transmitted diseases and infections depend upon numerous implications, each of which must be taken into account. While studying the epidemiological patterns of these conditions, one must individually analyze the interaction potential between host and pathogen, whether viral or bacterial, as well as interventions regarding health-care, before analyzing the potential causative associations where the pathogen may have been acquired. Since pathogen acquisition may hold the answer to interventions and preventative measures in the future, the use of social networking is a practice which may save much needed time and resources.

2. Generating realistic social networks of intimate interactions

We have developed a simulator capable of building a social network of heterosexual intimate contacts. This type of social network can be viewed as a bipartite graph described by the triplet $G = (G_f, G_m, E)$ where G_M represents male vertices, G_F represents female vertices and $E \subseteq G_f \times G_m$ is the set of edges. We first describe the general algorithm for our social network of intimate contacts simulator. Next, we define in detail how we perform bipartite matching on our network so that a minimum open degree remains. The authors implemented the simulator in C++ using the boost stl graph libraries [15].

2.1. General Algorithm

The general algorithm of our simulator contains several basic steps (Alg. 1). First, a forest is generated for each bipartition subset; next, the social network is created by linking the two subsets with each other based on several properties. The generated social network can then be used to evaluate disease dynamics and any intervention strategies.

begin
input : user defined parameter space
Insert n_m vertices in G_m
Insert n_f vertices in G_f
foreach $v_f \in G_f$ **do**
$maxDegree_{v_f} = zipf(-2.54, kFemaleBound)$
foreach $v_m \in G_m$ **do**
$maxDegree_{v_m} = zipf(-2.31, kMaleBound)$
$\forall v_f \in G_f$ calculate preferential attachment $p(k)$ $\forall v_m \in G_m$ calculate preferential attachment $p(k)$ **foreach** *Time Step* **do**
Generate bipartite network with minimum degree remaining
Evaluate disease dynamics including any intervention strategies
$\forall e_{u,v} \in G$ remove edge $e_{u,v}$
foreach $v_n \in G$ **do**
draw a uniformly distributed random number **r**
if $r > p(aging - out)$ **then**
remove v_n from G
create new node v_{new}
$maxDegree_{vnew} = zipf(\alpha, genderBound)$
insert v_{new} in G_k
recalculate preferential attachment $p(k)$ $forall$ $v_k \in G'$
end

Algorithm 1: General simulator algorithm

A forest is generated for each bipartition subset (G_m, G_f) in the graph G by inserting the respective number of nodes (n_m, n_f) specified by the user parameter space. A key factor that separates our networks from other affiliation networks is that of a specific degree distribution assigned to each subset. A Swedish survey on sexual behavior was analyzed and reported by Liljeros et al. in a 2001 Nature article [11, 12]. The survey was evaluated from a random sample of 4,781 Swedes ages 18-74 that involved questions and personal interviews. One of the survey questions was how many intimate partner changes occurred in a years time. Using the data obtained from this question, Liljeros et. al were able to determine a specific probability distribution for having k intimate partners. Males in the study reported a higher partner change rate than females; however, they both had similar scaling. In particular the paper cited the number of partners in the previous year follows a power law distribution. The cumulative probability function ($p(k) \approx L(k)k^{-\alpha}$) of a power-law distribution P_k is the probability of having k partners with scaling parameter $\alpha > 1$ and $L(k)$ being a slowly varying function that controls the shape and finite extent of the lower tail [13]. In our algorithm we use a specific type of power-law called a bounded Zipf-law, the authors chose this law so an exact upper bound (shown in [12]) could be placed on the number of intimate partner changes [18].

Random network models assume that a link may be placed randomly between two vertices and uniformly throughout the network. This is not the case in real world networks, where links are more likely to exist with non-random attachment. Preferential attachment results when a new node is more likely to connect to a node with a high degree than to a node of low degree. The probability of connecting to vertex ,v_j, Π_j is the connectivity of vertex v_j averaged over the total sum of each vertices degree ($\Pi_j = \frac{d_j^o}{\sum_{v_i \in G_k} d_{v_i}^o}$) [2, 3].

Once the population has been generated and preferential attachment probabilities have been assigned to each node. The time-driven simulation can commence. The first step is to maximally connect the two bipartition subsets to form our network of intimate contacts. The problem of finding the graph configuration with the lowest total remaining degree is a computationally intensive problem with a running time known to be in NP and represents an interesting dilemma. A heuristic described in section 2.2 is implemented to reduce the computation to $O(E \log V)$. After the contacts have been placed, disease dynamics and any intervention strategies can be performed on the network. The model maintains a constant population and it accounts for persons aging-out of the modeled age-span(υ = age span modeled). Nodes are stochastically removed based upon the probability of aging-out of the network ($\frac{1}{\upsilon}$). The removed nodes are

then replaced with new nodes, each new node is assigned a degree from the bounded Zipf-law distribution. Preferential attachment probabilities are recalculated and the network is then rebuilt according to algorithm described in Alg.1.

2.2. Maximally connecting a bipartite-graph

input : $G = (G_f, G_m, E)$ where $E = \emptyset$
output: Maximally connected bipartite graph of intimate contacts
begin
  while $\exists v_m \in G_m \; and \; \exists v_f \in G_f \; s.t. \; d^o_{v_k} < maxDegree_{v_k}$ **do**
    in vertex order given $v_{k_{id}}$
    choose $v_m \in G_m \; s.t. \; d^o_{v_m} < maxDegree_{v_m}$
    loopCount = 0
    maxReached = FALSE
    repeat
      inserted = FALSE
      loopCount++
      randomly choose $v_f \in G_f \; s.t. \; d^o_{v_f} < maxDegree_{v_f}$
      $p_E(Attach) = p_{v_f}(k) \times p_{v_m, v_f}(mixing)$
      draw a uniformly-distributed random number **r**
      if $p_E(Attach) >$ ***r*** **then**
        add $E = (v_f, v_m) \; in \; G$
        inserted = TRUE
      if *!inserted and maxLoopsReached* **then**
        arbitrarily choose $v_f \in G_f \; s.t. \; d^o_{v_f} < maxDegree_{v_f}$
        add $E = (v_f, v_m) \; in \; G$
        inserted = TRUE

    until *inserted == TRUE*
    in vertex order given $v_{k_{id}}$
    choose $v_f \in G_f \; s.t. \; d^o_{v_f} < maxDegree_{v_f}$
    loopCount = 0
    maxReached = FALSE
    repeat
      inserted = FALSE
      loopCount++
      randomly choose $v_m \in G_m \; s.t. \; d^o_{v_m} < maxDegree_{v_m}$
      $p_E(Attach) = p_{v_m}(k) \times p_{v_m, v_f}(mixing)$
      draw a uniformly-distributed random number **r**
      if $p_E(Attach) >$ ***r*** **then**
        add $E = (v_f, v_m) \; in \; G$
        inserted = TRUE
      if *!inserted and maxLoopsReached* **then**
        arbitrarily choose $v_m \in G_m \; s.t. \; d^o_{v_m} < maxDegree_{v_m}$
        add $E = (v_f, v_m) \; in \; G$
        inserted = TRUE

    until *inserted == TRUE*

end

Algorithm 2: Connecting a bipartite graph with a minimal number of remaining degree.

Let G be an undirected bipartite graph, that contains two bipartition subsets, G_M and G_F. The vertices in each subset have a pre-assigned degree associated with it; specifically, a random, power-law distributed number which is the maximum possible number of edges connected that vertex. For the constraint that each node is to have at least one edge the following bounds must hold $|N(G_f)| > |G_m|$ or $|N(G_m)| \geq |G_f|$. Note, the distribution of male and female vertices is not significant if the previously mentioned constraint holds. Edges are attached to the graph as follows: in vertex order of each subset, one edge is attached from a male vertice to a randomly chosen female vertice in the opposing subset, link attachment is determined by preferential attachment and demographic mixing probabilities. A threshold is set to arbitrarily choose a vertex in the opposite subset when a large number of vertices have been chosen but no edge has been placed. The simulator's threshold is set to 200 attempts before an arbitrary vertex in the opposite subset is chosen for edge placement. When a node randomly chooses a node in the opposite subset and it stochastically fails to create a link, the model will draw a random node 200 times before arbitrarily choosing a vertex for edge placement; the number 200 is arbitrarily chosen and sensitivity on this threshold is left for future work. Next, an edge is attached from a female vertex to a male vertex, also determined by preferential attachment and demographic mixing probabilities. The edges are added one per subset until one of the subsets maximum cardinality is reached. The exact algorithm is described in greater detail in Alg. 2.

3. Bipartite graph statistics

Analyzing graphs with various statistical properties have become an important component in describing real world complex networks. Many of the bipartite graph statistics relate to their classical counterparts. Clustering coefficients in classical graphs measure overlap with triangles; however, triangles between vertices do not occur in bipartite graphs. A recent paper by Latapy et al. describes the following bipartite graph statistics in greater detail [10]. First, we define a neighborhood $N(v)$ of vertex v as $N(v) = \{\{u\} : e_{u,v} \in E\}$. Equations 1-3 calculate the clustering coefficient for a pair of vertices, a single vertex, and the graph. Equations 4-5 define the bounds of clustering coefficients. The following clustering coefficients can be evaluated similarly to Eqns. 2 and 3 : $cc_{\downarrow}(u)$, $cc_{\downarrow}(\top)$, $cc_{\downarrow}(\bot)$, $cc_{\downarrow}(G)$, $cc_{\uparrow}(u)$, $cc_{\uparrow}(\top)$, $cc_{\uparrow}(\bot)$, and $cc_{\uparrow}(G)$[10].

$$cc_{\bullet}(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{1}$$

$$cc_{\bullet}(u) = \frac{\sum_{v \in N(N(u))} cc_{\bullet}(u, v)}{|N(N(u))|} \tag{2}$$

$$cc_{\bullet}(G) = \frac{n_{\top} cc_{\bullet}(\top) + n_{\bot} cc_{\bullet}(\bot)}{n_{\top} + n_{\bot}} \tag{3}$$

$$cc_{\uparrow}(u, v) = \frac{|N(u) \cap N(v)|}{\min(|N(u)|, |N(v)|)} \tag{4}$$

$$cc_{\downarrow}(u, v) = \frac{|N(u) \cap N(v)|}{\max(|N(u)|, |N(v)|)} \tag{5}$$

4. Experimental Results

We have introduced several methods to examine the topology of complex bipartite networks. Next, we evaluate the graphs produced by our social network generator using Monte-Carlo type simulations. The computational complexity associated with calculating bipartite graph statistics allowed for ten runs, each with ten time steps. Each run generates a social network of intimate contacts and each time step redistributes contacts throughout the population;

graph statistics are calculated at the end of each time step. The networks contain 10,000 vertices and each bipartition subset has an equal number of male and female vertices ($|G_m| = |G_f|$).

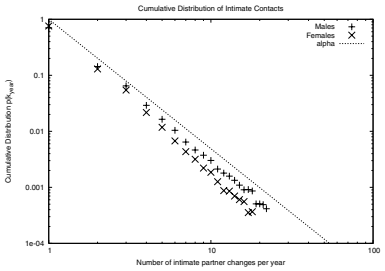

Figure 1. Cum. Dist. of Intimate Contacts

The cumulative distribution for intimate contacts is displayed on a $\log - \log$ plot in Figure 1. The solid line demonstrates a power law curve with $\alpha = 2.31$, it can be seen that the contacts generated by our simulator slightly under-fit the original distribution however the scaling remains. Currently, our model slightly under-fits the power-law scaling reported by Liljeros et al.; this is due to when a node reaches its maximum degree we do not allow (by chance) for a link to be added to that vertex [12]. Note that approximately 90% of the vertices have only one contact (shown in Fig.1) and thus result in approximately 4,500 links; the average edge count for our networks is 7693 and 10% of the vertices account for $\approx$ 3200 links. The graph statistics resulting from the Monte-Carlo simulations are displayed in Table 1. They show the range of clustering coefficients for the graph and each bipartition subset. Each specific clustering coefficient statistic show a high occurance of clusters compared to the density of G and G'. Currently, only statistics are calculated from networks generated solely by preferential attachment. Future work will include comparison between preferential attachment combined with the demographic mixing probabilities and preferential attachment alone.

Statistic	Result
m	7693
$\delta(G)$	2.90E-4
$\delta(G')$	1.45E-4
$cc(Males)$	0.423
$cc(Females)$	0.218
$cc(G)$	0.321
$cc_{\uparrow}(Males)$	0.786
$cc_{\uparrow}(Females)$	0.848
$cc_{\uparrow}(G)$	0.817
$cc_{\downarrow}(Males)$	0.456
$cc_{\downarrow}(Females)$	0.226
$cc_{\downarrow}(G)$	0.321

Table 1. Generated SN Statistics

Many Human Papilloma Virus (HPV) types are sexually transmitted and HPV DNA is found in 99.7% of all cervical cancers with HPV-types 16, 18, 31 and 45 accounting for 75% of cervical dysplasia[8]. HPV prevalence is a large component of cervical cancer's etiology which we use to demonstrate the capability of the simulator. We evaluate the impact of several disparate intervention strategies on HPV prevalence in the population. The simulator's parameter space is gathered from [4]. To determine the probability of natural infection a binomial is calculated with the chance of infection in one encounter (p_{i_k}) and the number of encounters (λ) which occur ($p_n(i) = 1 - [1 - p_{i_k}]^{\lambda}$); similarly, the probability of breakthrough infection combines intervention efficacy (e_{int}) and chance of natural infection ($p_b(i) = e_{int} \times p_{i_k}$). The specific stochastic disease parameters include the probability of acquiring HPV in one encounter (0.08 male-to-female, 0.02 female-to-male), encounter frequency drawn from a Poisson distribution with a mean of 50, intervention efficacy is 75%, the age-range modeled is 50 years, infection clears after two years and 5% of the population is initially infected.

Population-level impact from three intervention strategies is evaluated; these include no intervention, vaccinating[1] only males, and vaccinating only females. An intervention targeting both males and females would be cost-prohibitive and not included in our evaluations. Each Monte-Carlo simulation is loaded with the parameter space described earlier, population size of 1,000 ($|G_m| = |G_f|$), and represents 30 years. The impact of each intervention setting is averaged from 100 Monte-Carlo simulations and the results are shown in Fig.2. Intervention results are analyzed by the relative reduction in prevalence (RRP) between no intervention and a specific strategy. Our results show a RRP of 75% (0.2 to 0.05) at the height of the epidemic. To date, no other social network simulator solely built on heterosexual intimate contacts has been developed for intervention analysis; however, much research has been conducted in this area using mean-field type and ordinary-differential equation models. Hughes et al cite a RRP of 0.68 with a range of 0.628 to 0.734; other models such as Sanders and Taira cite a RRP of 0.8 and above[9, 14]. Endemic prevalence does not occur with our simulator[2]; however, our results clearly show a reduction in prevalence within the RRP range of established models.

5. Conclusions

Recent growth in the prevalence of sexually transmitted diseases and infections in developing and developed countries general population has prompted a great deal of inter-disciplinary research to curb the population wide effect of these diseases. Public health professionals often have lim-

[1] Intervention coverage is 80%.

[2] This is due to how the network is currently generated using coarse attachment probabilities

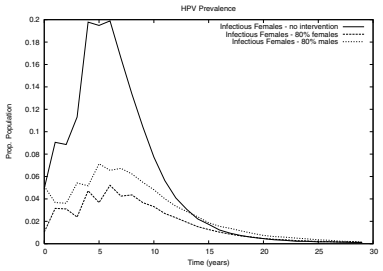

Figure 2. Population-level impact of HPV interventions

ited budgets and resources must be specifically tailored to achieve maximum results. The utilization of computational social networking tools would allow for those within the public health industry to anticipate the impact of demographic specific predictions, and tailor awareness, educational, vaccination, and prophylactic programs for the greatest impact within their population.

Our social network generator is in the foundation phase of development and there is exciting future work to be accomplished. We analyzed the current networks which are generated by using only preferential attachment as the contact likelihood. The next phase of development will assign social demographic properties to each node and combine preferential attachment with the likelihood of mixing between these social demographic groups. Evaluating several different contact placement options will lead to a more precise social network generated. Examples of these contact placement strategies include placing edges by randomly choosing a node from each bipartition subset and stochastically choosing placement, exhausting a single nodes total degree before iterating to the next node and exhausting only one bipartition subsets total degree. The correctness proofs of the contact placement heuristics are beyond the scope of this paper.

We introduced a novel algorithm to generate social networks of intimate contacts. The general algorithm generates a contact driven network with specific degree distribution and a dynamic population. Next a simple heuristic was introduced capable of performing bipartite matching in polynomial time reducing the computation power needed for the simulation from NP to $E \log V$. Several graph-analytic methodologies were introduced that facilitate evaluation of the generated social networks; in particular, bipartite graph statistics. Disease dynamics can then be analyzed on the generated networks along with tailored intervention strategies to provide what-if analyses.

6. Acknowledgements

We would like to thank the National Science Foundation for support under grant NSF IIS-0505819 and the authors of the boost graph libraries (www.boost.org) for the use of their C++ stl graph packages in our simulator. This publication was also made possible by Grant Number P20-MD001633 from NCMHD, its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCMHD.

References

[1] S. Aral, J. Hughes, B. Stoner, W. Whittington, H. Handsfield, R. Anderson, and K. Holmes. Sexual mixing patterns in the spread of gonococcal and chlamydial infections. *American Journal of Public Health*, 89(6):825–833, 1999.

[2] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] A. Barabasi and R. Albert. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[4] C. Corley and A. Mikler. Predicting human papilloma virus prevalence and vaccine policy effectiveness in demographic strata. In *Proceedings of IEEE fifth symposium on bioinformatics and bioengineering (BIBE05)*, Minneapolis, MN, October 2005.

[5] K. Eames and M. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc Natl Acad Science*, 99(20):13330–13335, 2002.

[6] T. Eng and W. Butler. *The hidden epidemic*. National academy press, 1996.

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

[8] S. Goldie, Kohli, and D. Grima. Projected Clinical Benefits and Cost-effectiveness of a Human Papillomavirus 16/18 Vaccine. *National Cancer Institute*, 96(8):604–615, April 2004.

[9] J. Hughes, G. Garnett, and L. Koutsky. The Theoretical Population-Level Impact of a Prophylactic Human Papilloma Virus Vaccine. *Epidemiology*, 13(6):631–639, November 2002.

[10] M. Latapy, C. Magnien, and N. Del Vecchio. Basic notions for the analysis of large affiliation networks / bipartite graphs. arXiv.org:cond-mat/0611631, 2006.

[11] B. Lewin. Sex in sweden. on the sexual life in sweden 1996. *Natl Inst Pub Health*, 1998.

[12] F. Liljeros, C. Edling, L. Amaral, H. Stanley, and Y. Aberg. Human web of sexual contacts. *Nature*, 411:907–908, 2001.

[13] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[14] G. Sanders and A. Taira. Cost Effectiveness of a Potential Vaccine for Human Papillomavirus. *Emerging Infectious Diseases*, 9(1):37–48, January 2003.

[15] J. Siek, L. Lee, and A. Lumsdaine. The boost graph libraries www.boost.org/libs/graph.

[16] H. Ward, C. Ison, S. Day, I. Martin, A. Ghani, G. Garnett, G. Bell, G. Kinghorn, and J. Weber. A prospective social and molecular investigation of gonococcal transmission. *Lancet*, 356:1812–1817, 2000.

[17] J. Wylie and A. Jolly. Patterns of chlamydia and gonorrhea infection in sexual networks in manitoba, canada. *Sexually Transmitted Diseases*, 28(1):14–24, 2001.

[18] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, 1949.