
Dynamic intimate contact social networks and epidemic interventions.

Courtney D. Corley¹, Armin R. Mikler¹,
Diane J. Cook², and Karan P. Singh³

¹Computational Epidemiology Research Laboratory, Department of Computer Science and Engineering, University of North Texas, Denton, TX, 76203 USA

²Department of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164 USA

³Department of Biostatistics, University of North Texas Health Science Center, Fort Worth, TX 76107 USA

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 dynamic and evolving social network simulation. This paper presents DynSNIC (Dynamic Social Network of Intimate Contacts), a computational simulator created to embody the intimate dynamic and evolving social networks related to the transmission of sexually transmitted diseases and infections. DynSNIC's utilization by health professionals will facilitate evaluation of targeted intervention strategies and public health policies.

Keywords: dynamic social network, epidemic interventions, public health, simulation, computational epidemiology

Reference to this paper should be made as follows: Corley, Mikler, Cook, and Singh (2008) 'Dynamic intimate contact social networks and epidemic interventions', *Int. J. Functional Informatics and Personalized Medicine*, Vol. 1, No. 2, pp.xxx-xxx.

Biographical notes: Mr. Courtney D. Corley is a doctoral candidate in the Department of Computer Science and Engineering at the University of North Texas. Mr. Corley received his B.S. and M.S. degrees in Computer Science from the University of North Texas in 2004 and 2006 respectively. His research interests include dynamic social networks, health informatics, mathematical modeling and natural language processing.

Armin R. Mikler received his Diploma in Informatics from Fachhochschule Darmstadt, Germany in 1988. After spending one year as a Fulbright exchange student at Iowa State University (ISU), he joined ISU as a graduate student and received his MS and Ph.D. in Computer Science in 1990 and 1995 respectively. In 1997, Dr. Mikler joined the faculty in Computer Science at the University of North Texas (UNT) where he holds the rank of associate professor in Computer Science with joint

appointment in the Department of Biological Sciences. His research interests include: Computational Epidemiology, Distributed Coordination of Intelligent Mobile Agents, Distributed Decision Making, Multi-Agent Simulation and Stochastic Cellular Automata.

Dr. Diane J. Cook is a Huie-Rogers Chair Professor in the School of Electrical Engineering and Computer Science at Washington State University. Dr. Cook received a B.S. degree in Math/Computer Science from Wheaton College in 1985, a M.S. degree in Computer Science from the University of Illinois in 1987, and a Ph.D. degree in Computer Science from the University of Illinois in 1990. Her research interests include artificial intelligence, machine learning, graph-based relational data mining, smart environments, and robotics.

Dr. Karan P. Singh is Professor and Chair in the School of Public Health Department of Biostatistics at University of North Texas Health Science Center at Fort Worth University. Dr. Singh received a B.Sc. degree in Mathematics, Physics and Chemistry from Meerut University (India) in 1975, a M.Sc. degree in Applied Statistics from the CCSHA University (India) in 1977, a M. S. degree in Computational and Applied Mathematics from the Old Dominion University in 1982, and a Ph.D. degree in Statistics from the University of Memphis in 1986. His research interests include mathematical modeling, longitudinal data analysis, statistical computing, survival methodology and health services research methodology.

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 (Eames & Keeling 2002). 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 (Eng & Butler 1996). 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.

Sexually transmitted diseases and infections are on the brink of becoming considered endemic within general populations. Many of these illnesses are preventable in nature, and the public health industry would benefit from the predictive measures capable of intimate social networking computational tools. Professionals within this field often have limited budgets and resources must be aimed in the proper direction in order 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.

With limited funding and resources available to help prevent infectious disease, public health professionals need tools to facilitate decision making regarding where

the most effective measures would be taken. Based on collected data and statistical analysis, it is evident that certain demographic groups are at higher risk for contracting certain sexually transmitted diseases and infections. For example, previous research has indicated specific achieved levels of education have a positive correlation with higher incidence of HIV/AIDS infection (Reiche & et al. 2005). This type of understanding, when applied to a computational social model, would allow the individual within the public health industry to model an awareness or educational campaign to the population with the greatest risk factors and to predict the potential impact on this target group from avoiding future infection.

In this paper we first introduce several methodologies to analyze graphs, in particular classical graphs and their mapping on to bipartite networks; for example: size, density, and clustering coefficients. The general algorithm of our dynamic social network of intimate contacts (DynSNIC) simulator is presented. This algorithm generates a dynamic contact driven network with a specific degree distribution, disease dynamics and evolving population. We then describe in detail how DynSNIC optimizes the bipartite network with a predetermined degree distribution, minimizing the number of unresolved degrees. The networks generated are then analyzed using the graph statistics introduced earlier in this paper. A sample case study is presented demonstrating DynSNIC's capabilities and this paper concludes with future work in our simulator's development.

2 Previous Work

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 (Eubank, Guclu, Kumar, Marathe, Srinivasan, Toroczkai & Wang 2004). This computational model may be used to handle diverse social networking in regard 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. Since its inception, EPISIMS research has relocated to the Virginia Bioinformatics Institute at Virginia Tech. Their research has expanded to include simulation of most cities, a coarser grained simulation of the entire U.S, and multiple versions of EPISIMS based on various modeling paradigms. The Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University has also built a simulation framework, BIOWAR, to predict the effect of a large-scale terrorist attack or infection outbreak. BIOWAR incorporates multi-agent systems, census track data, human social behavior and wind dispersion data (Carley, Altman, Casman, Fridsma, Yahja, Chen, Kaminsky & Nave 2006).

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. Gonorrheal and Chlamydial infections have been predicted using these types of models, which have been validated through available statistics. Due to the nature of these illnesses, statistical data is often

very difficult to collect (Aral, Hughes, Stoner, Whittington, Handsfield, Anderson & Holmes 1999, Ward, Ison, Day, Martin, Ghani, Garnett, Bell, Kinghorn & Weber 2000, Wylie & Jolly 2001). Epidemiological models for sexually transmitted conditions have also been created based on the accumulation of contact tracing data. This type of data may be unreliable due to individual recall error and privacy constraints, but is the common method of understanding Syphilis transmission (Chen, Kodagoda, Lawrence & Kerndt 2002, Gunn, Harper, Borntrager, Gonzales & St.Louis 2000, Rosenberg, Moseley, Kahn, Kissinger, Rice, Kendall, Coughlin & Farley 1999, Williams, Klausner, Whittington, Handsfield, Celum & Holmes 1999, Wylie & Jolly 2001).

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.

3 Graph statistics

3.1 Classical graph statistics

Analyzing graphs with various statistical properties has become an important component in describing real world complex networks. First, we briefly introduce basic graph-theoretic statistics including clustering coefficients. Next, we describe the modification of these methods to analyze properties of bipartite graphs..

The analysis of classical graphs is a well studied field in graph-theory and many methodologies exist to describe the nature of these graphs. Traditionally, a classical graph, G , is defined $G = (V, E)$ where V is the set of vertices and E is the set of edges in the graph $E \subseteq V \times V$. The neighborhood $N(v)$ of vertex v is defined as $N(v) = \{\{u\} : e_{u,v} \in E\}$. The degree of vertex v is the cardinality of the set of edge connections from v to its neighborhood, $d_v^o = |N(v)|$. Basic statistics that describe this graph include its size $n = |V|$, number of edges $m = |E|$, average degree $k = \frac{2m}{n}$, and its density, $\delta(G)$, which represents the probability any two randomly chosen vertices are connected, $\delta(G) = \frac{2m}{n(n-1)}$.

In addition, we consider two more statistics in the context of graphs, degree distribution $p(k)$ and clustering coefficients. Degree distribution gives the probability of degrees in a network and has become an integral descriptive of the topology of complex networks. The degree distribution function $p(k)$ describes the total number of vertices in a graph with a given degree. This same information is also described by the cumulative degree distribution (Eq. 1) (Erdos & Renyi 1959).

$$P_k = \sum_{k'=k} p_{k'} \tag{1}$$

The second graph descriptor is the clustering coefficient. It has been empirically shown that many social networks have a higher neighborhood transitivity than that

of other random networks such as Internet topology (Newman & Park 2003, Watts & Strogatz 393). Much of the analysis of the networks generated by our simulator is evaluated using clustering coefficients. This statistic describes the overlap in the network topology. The clustering coefficient C_v is the probability that any two nodes are linked together if they have a neighbor in common. In an undirected graph $e(u, v)$ and $e(v, u)$ are the same link. Hence, if vertex v has k neighbors $\frac{k(k-1)}{2}$ edges could exist in the neighborhood. Equation 2 defines the clustering coefficient for undirected graphs. The clustering coefficient for the entire graph is the averaged sum of vertices' clustering coefficients. (Eq. 3) (Watts & Strogatz 393).

$$C_{\bullet v} = \frac{|E_{N(v)}|}{\frac{|N(v)|(|N(v)|-1)}{2}} = \frac{2|\{e(y, u)\}|}{d_v^o(d_v^o - 1)} \quad (2)$$

$$: y, u \in N(v), e(y, u) \in E$$

$$\bar{c}_{\bullet} = \frac{1}{n} \sum_{i=1}^n C_{\bullet i} \quad (3)$$

3.2 Bipartite graph statistics

Many of the bipartite graph statistics relate to their classical counterparts. Some of these descriptors are redefined while others are dual components of their classical property. A recent technical paper by Latapy et al. describes the following bipartite graph statistics in greater detail (Latapy, Magnien & Del Vecchio 2006).

Consider a bipartite graph $G = (\top, \perp, E)$. The size of the graph is now divided into the size of the top portion $n_{\top} = |\top|$ and the size of the bottom subset $n_{\perp} = |\perp|$, these are the number of nodes in the top vertex set and the bottom set, respectively. The size of the edge set remains the same as for classical graphs $m = |E|$. The average degree is now separated for each bipartition subset; the top subsets average degree is $k_{\top} = \frac{m}{n_{\top}}$ and the bottom subset $k_{\perp} = \frac{m}{n_{\perp}}$. The average degree of the graph $G^* = (\top \cup \perp, E)$ is now $k = \frac{2m}{n_{\top} + n_{\perp}} = \frac{n_{\top} k_{\top} + n_{\perp} k_{\perp}}{n_{\top} + n_{\perp}}$. The bipartite density is thus $\delta(G) = \frac{m}{n_{\top} n_{\perp}}$ and $\delta(G^*) = \frac{2m}{(n_{\top} + n_{\perp})(n_{\top} + n_{\perp} - 1)}$ with $\delta(G^*) \ll \delta(G)$.

Clustering coefficients are evaluated much differently in the bipartite setting. In the classical graphs, the overlap among vertices is measured by the number of triangles; however, in the bipartite case, triangles among vertices of the same set do not occur. The following descriptors will be used to analyze the topology of the networks generated by our simulator. We define the clustering coefficient for pairs of nodes, both in either \top or \perp : cc_{\bullet} captures the overlap in neighborhoods of vertices u and v . Whenever the neighborhood of vertices u and v do not overlap then $cc_{\bullet}(u, v) = 0$. Conversely, if vertices u and v are elements of the same neighborhood then $cc_{\bullet}(u, v) = 1$. The equation for the neighborhood overlap is given in Eq. 4. The cartoon in Fig.1 demonstrates clustering in a bipartite graph. The neighborhoods of vertices u and v intersect at nodes a and b in the opposing subset, the corresponding clustering coefficient is $cc_{\bullet}(u, v) = \frac{2}{3}$; however, there is no neighborhood intersection between vertices u and w and $cc_{\bullet}(u, w) = 0$. To evaluate the clustering coefficient of a particular node, the average over the subset is calculated for only those edge pairs where an overlap in neighborhoods exist

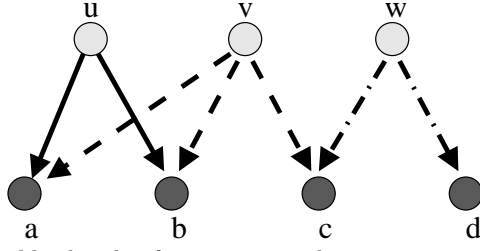


Figure 1 The neighborhoods of vertices u and v intersect at vertices a and b . The clustering coefficient between these two vertices is $cc_{\bullet}(u, v) = \frac{2}{3}$. However, there is no overlap (clustering) between vertices u and w ; thus the clustering coefficient of vertex u remains the same ($cc_{\bullet}(u) = \frac{2}{3}$).

(Eq. 5). The graphs clustering coefficient $cc_{\bullet}(G)$ is the average of each bipartition subsets corresponding clustering coefficient ($cc_{\bullet}(\top), cc_{\bullet}(\perp)$) (Eq. 6). Considering complex networks with significant differences between degrees of the vertices, the previously introduced clustering coefficient may not provide a strong and informative analysis of the network topology. The following two clustering coefficient *flavors* further describe neighborhood overlap. Equation 7 describes a clustering coefficient lower bound and considers a setting where a small neighborhood is encompassed by a large neighborhood. Equation 8 evaluates an upper bound on the clustering coefficient and evaluates occurrences where small or large neighborhoods overlap. The following clustering coefficients can be evaluated similarly to Eq. 5 and 6: $cc_{\downarrow}(u)$, $cc_{\downarrow}(\top)$, $cc_{\downarrow}(\perp)$, $cc_{\downarrow}(G)$, $cc_{\uparrow}(u)$, $cc_{\uparrow}(\top)$, $cc_{\uparrow}(\perp)$, and $cc_{\uparrow}(G)$ (Latapy et al. 2006).

$$cc_{\bullet}(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \quad (4)$$

$$cc_{\bullet}(u) = \frac{\sum_{v \in N(N(u))} cc_{\bullet}(u, v)}{|N(N(u))|} \quad (5)$$

$$cc_{\bullet}(G) = \frac{n_{\top} cc_{\bullet}(\top) + n_{\perp} cc_{\bullet}(\perp)}{n_{\top} + n_{\perp}} \quad (6)$$

$$cc_{\uparrow}(u, v) = \frac{|N(u) \cap N(v)|}{\min(|N(u)|, |N(v)|)} \quad (7)$$

$$cc_{\downarrow}(u, v) = \frac{|N(u) \cap N(v)|}{\max(|N(u)|, |N(v)|)} \quad (8)$$

4 Generating realistic social networks of intimate interactions

We have developed a simulator capable of building an evolving social network of dynamic heterosexual intimate contacts. This type of social network can be viewed as a bipartite graph described by the triplet $G = (G_f, G_m, \vec{E})$ where G_m represents male, G_f represents female vertices contained in the network at any

point in time and $\{\vec{\mathbf{E}} : \vec{\mathbf{E}} \subseteq G_f \times G_m\}$ is a vector containing the set of edges present during discrete time intervals. 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 (Siek, Lee & Lumsdaine 2008).

4.1 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.

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. Demographic characteristics play an important role in intimate interaction among individuals. To encapsulate these characteristics, each vertex is assigned a vector of demographic properties $(\vec{\mathbf{D}}_{v_k})$. Each vector component is labeled by a specific feature set and its value is arbitrarily chosen from the set's range of discrete values. Due to limited available data, we choose not to explicitly identify each feature (i.e. race, income, education) and simply assign each vector component a probability distribution. Currently, gender is treated as a special case and not included in the feature vector; this is due to only heterosexual interactions consideration in the network.

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 (Eq. 9) (Barabási & Albert 1999, Barabási & Albert 2002).

$$\Pi_j = \frac{d_j^o}{\sum_{v_i \in G_k} d_{v_i}^o} \quad (9)$$

Demographic similarity can either strengthen or weaken the probability of connection, considering assortative to random mixing. Many scoring metrics exist to quantify similarity between two objects; for example, hamming distance, cosine similarity and feature frequency proportions (Hamming 1950, Jaccard 1901, Salton & McGill 1983). DynSNIC takes a coarse first cut at scoring the likelihood of mixing ($p_{v_m, v_f}(\text{mixing})$) between two individuals, using the unweighted cosine similarity of both vertices's demographic feature vector (Eq. 10). When selecting an intimate partner in actual social settings, it is likely that certain demographic features provide stronger, weaker, or even negative attraction (i.e. education, income, age), which would lead to a weighted function. Currently, the simplest (unweighted) *flavor* of the similarity scoring metrics is implemented in DynSNIC's initial experiments. The overall likelihood of an interaction occurring ($p_E(\text{Attach})$) between

```

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
     $maxDegree_{v_f} \leftarrow zipf(-2.54, kFemaleBound)$ 
    draw a random vector  $\tilde{\mathbf{r}}$  from a user defined and bounded probability
    distribution
     $\vec{\mathbf{D}}_{\mathbf{v}_f} \leftarrow \tilde{\mathbf{r}}$  //demographic property vector stochastically created and
    assigned to female vertice
  foreach  $v_m \in G_m$  do
     $maxDegree_{v_m} \leftarrow zipf(-2.31, kMaleBound)$ 
    draw a random vector  $\tilde{\mathbf{r}}$  from a user defined and bounded probability
    distribution
     $\vec{\mathbf{D}}_{\mathbf{v}_m} \leftarrow \tilde{\mathbf{r}}$  //demographic property vector stochastically created and
    assigned to male vertice
   $\forall v_f \in G_f$  calculate preferential attachment  $p(k)$ 
   $\forall v_m \in G_m$  calculate preferential attachment  $p(k)$ 
  foreach discrete time step do
     $E_t = \{\emptyset\}$  //initialize current edge set to the empty set
     $optBiMatching(G')$  //Optimize bipartite matching on dynamic network
    , see Alg.2
    Evaluate disease dynamics including any intervention strategies
    foreach  $v_n \in G'$  do
      //population evolution
      draw a uniformly distributed random number  $r$ 
      if  $r > p(aging - out)$  then
        color  $v_n \in G$  as unusable
        create new node  $v_{new}$ 
         $maxDegree_{v_{new}} \leftarrow zipf(\alpha, genderBound)$ 
        insert  $v_{new}$  in  $G_k$ 
        recalculate preferential attachment  $p(k) \forall v_k \in G'$ 
    end
Algorithm 1: DynSNIC general algorithm. Note :  $G' = \{G_{m_{usable}}, G_{f_{usable}}, E_t\}$ 

```

two vertices (v_f and v_m) is the aggregate of the scoring function and preferential attachment ($p_E(Attach) = p_{v_m}(k) \times p_{v_m, v_f}(mixing)$).

$$p_{v_m, v_f}(mixing) = \text{COSIM}(\vec{\mathbf{D}}_{\mathbf{v}_f}, \vec{\mathbf{D}}_{\mathbf{v}_m}) = \cos \theta = \frac{\vec{\mathbf{D}}_{\mathbf{v}_f} \cdot \vec{\mathbf{D}}_{\mathbf{v}_m}}{|\vec{\mathbf{D}}_{\mathbf{v}_f}| |\vec{\mathbf{D}}_{\mathbf{v}_m}|} \quad (10)$$

A key factor in assessing structure among intimate connections, as is common with most other social network topologies, is degree distribution. A Swedish survey on sexual behavior was analyzed and reported by Liljeros et al. in a 2001 Nature article (Lewin 1998, Liljeros, Edling, Amaral, Stanley & Aberg 2001). 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 (Eq. 11) 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 (Newman 2005). 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 (Liljeros et al. 2001)) could be placed on the number of intimate partner changes (Zipf 1949).

$$p(k) \approx L(k)k^{-\alpha} \tag{11}$$

Once the population has been generated, preferential attachment probabilities and demographic feature vectors have been assigned to each node; the time-driven simulation can commence. The first step is to maximally connect the two bipartition subsets forming a network of intimate interactions. 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 *NP-hard* and represents an interesting dilemma. A greedy-heuristic described in section 4.2 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's population size remains constant; however, it evolves through accounting for persons aging-out of the modeled age-span ($v = \text{age span modeled}$). Nodes are stochastically colored unusable based upon the probability of aging-out of the network ($\frac{1}{v}$). The unusable nodes are then replaced with new nodes, each new node is assigned a demographic feature vector, and its cardinality from the bounded Zipf-law distribution. Preferential attachment probabilities are recalculated for each vertex and the dynamic network is then rebuilt according to algorithm described in Alg.1.

4.2 Maximally connecting a bipartite-graph : *optBiMatching*(G')

The purpose of Alg. 2 (*optBiMatching*(G')) is to connect every node in G_m to another node in G_f , maximally exhausting each subsets total degree. Algorithm 2 optimizes matching on the bipartite graph in polynomial time ($O(E \log V)$) with the following constraints : every node has at least one connected edge and the resulting graph's cardinality is optimized so that a minimal number of edges remain to be connected.

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

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. $d_{v_k}^o < \maxDegree_{v_k}$ **do**
 //in vertex order given $v_{k_{id}}$
 choose $v_m \in G_m$ s.t. $d_{v_m}^o < \maxDegree_{v_m}$
 loopCount = 0
 maxReached = FALSE
 repeat
 inserted = FALSE
 loopCount++
 randomly choose $v_f \in G_f$ s.t. $d_{v_f}^o < \maxDegree_{v_f}$
 $p_E(Attach) = p_{v_f}(k) \times p_{v_m, v_f}(mixing)$
 draw a uniformly-distributed random number \mathbf{r}
 if $p_E(Attach) > \mathbf{r}$ **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. $d_{v_f}^o < \maxDegree_{v_f}$
 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. $d_{v_f}^o < \maxDegree_{v_f}$
 loopCount = 0
 maxReached = FALSE
 repeat
 inserted = FALSE
 loopCount++
 randomly choose $v_m \in G_m$ s.t. $d_{v_m}^o < \maxDegree_{v_m}$
 $p_E(Attach) = p_{v_m}(k) \times p_{v_m, v_f}(mixing)$
 draw a uniformly-distributed random number \mathbf{r}
 if $p_E(Attach) > \mathbf{r}$ **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. $d_{v_m}^o < \maxDegree_{v_m}$
 add $E = (v_f, v_m)$ in G
 inserted = TRUE
 until inserted == TRUE
end

Algorithm 2 $optBiMatching(G')$ where $G' = \{G_{m_{usable}}, G_{f_{usable}}, E_t\}$. Connecting a bipartite graph minimizing remaining degree.

is determined by preferential attachment and cosine similarity of each vertices’s demographic vector. 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 cosine similarity of each vertices’s demographic vector. 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. A sample network generated by our simulator that contains 100 nodes, 50 females and 50 males is displayed in Fig. 2.

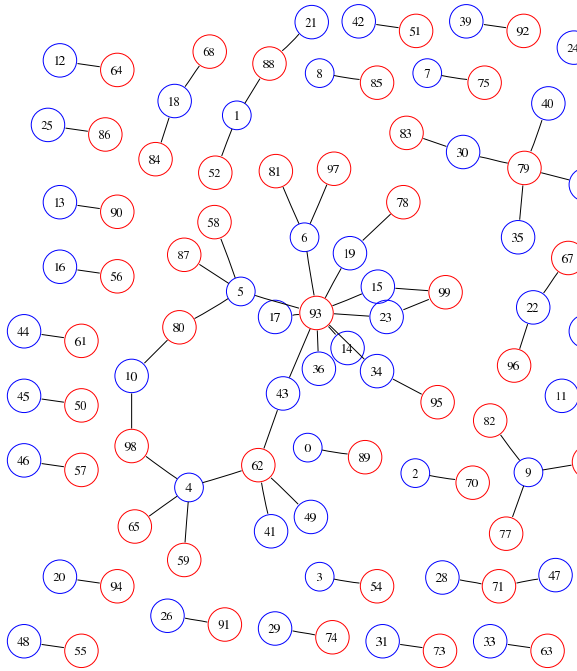


Figure 2 Portion of a 100 node ($|G_m| = |G_f|$) social network realization where the probability of intimate connection is $p(k) \times \text{COSIM}(\mathbf{D}_{v_m}^{\rightarrow}, \mathbf{D}_{v_f}^{\rightarrow})$.

5 Experimental Results

We have introduced several methods to examine the topology of complex bipartite networks. Next, we evaluate the dynamic networks produced by DynSNIC using Monte-Carlo type simulations. The computational complexity associated with calculating bipartite graph statistics allowed for ten runs, each run generating ten contact realizations of the dynamic social network. The networks contain

10,000 vertices and each bipartition subset has an equal number of male and female vertices ($|G_m| = |G_f|$).

The partner change cumulative distribution is displayed on a log – log plot in Fig. 3. 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 (Liljeros et al. 2001). Note that approximately 90% of the vertices have only one contact (shown in Fig. 3) and thus result in approximately 4,500 links; the average edge count for our preferential attachment networks is 7693 and 10% of the vertices account for ≈ 3200 links (similar statistics are present for cosine similarity and PAXCOSIM networks).

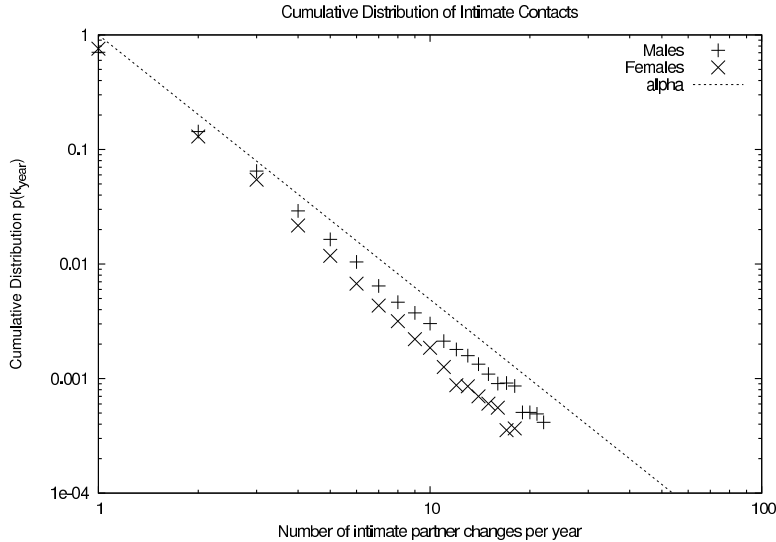


Figure 3 Cumulative Distribution for Intimate Contacts

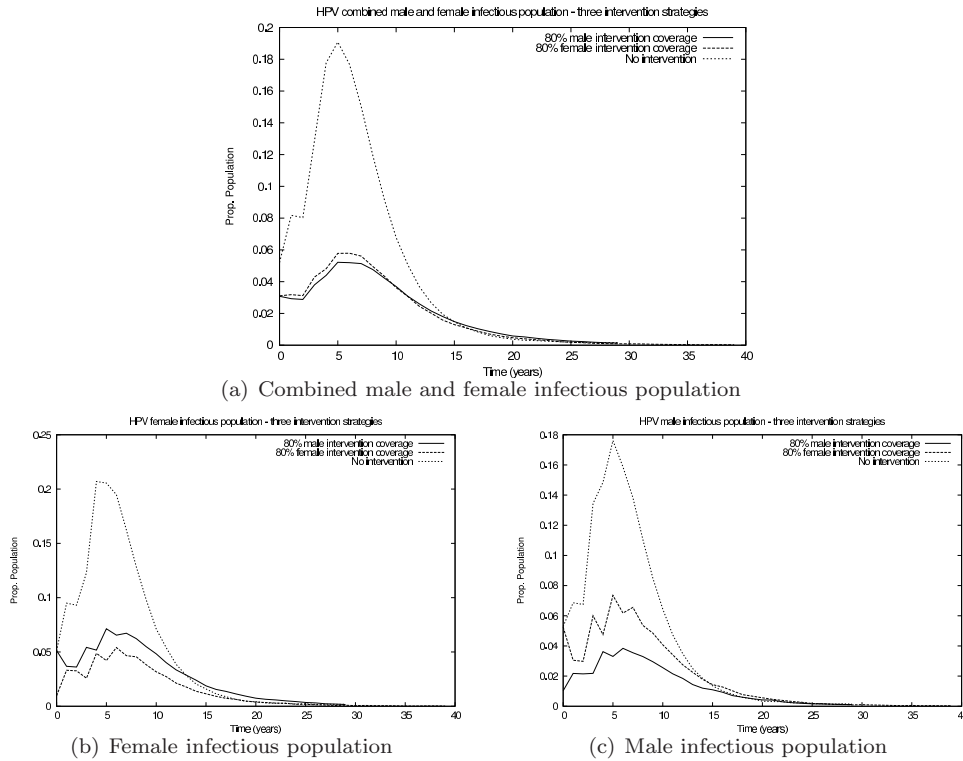
Three evaluation settings were chosen for our experiments, defined by interaction probabilities. The interaction likelihood settings are using solely demographic feature vector cosine similarity, solely preferential attachment, and the aggregate of preferential attachment and cosine similarity scoring. The graph statistics resulting from the Monte-Carlo simulations are displayed in Table 1. The results demonstrate the likelihood of clusters in our networks. They also show the range of clustering coefficients for the graph and each bipartition subset. Each specific clustering coefficient statistic show a high occurrence of clusters compared to the density of G and G^* . An interesting observation from the table is that preferential attachment alone, produces less neighborhood overlap compared to demographic feature cosine similarity. This correlates to results from comparing clustering in social networks to that of non-human networks where interactions are defined more by network topology than other affinity measures; for instance, the Internet topology compared to the Live-Journal community (Kumar, Novak, Raghavan & Tomkins 2004). Centrality measures are valuable in quantifying network topologies; evaluating these

metrics on dynamic and evolving networks is an open research question and the authors leave centrality evaluation to future work (Berger-Wolf & Saia 2006).

Statistic	$p(k)$	COSIM	$p(k) \times \text{COSIM}$
m	7693	7716	7742
$\delta(G)$	3.08E-4	3.09E-4	3.10E-4
$\delta(G^*)$	1.54E-4	1.54E-4	1.55E-4
$cc(\text{Males})$	0.423	0.457	0.442
$cc(\text{Females})$	0.218	0.250	0.233
$cc(G)$	0.321	0.353	0.337
$cc_{\uparrow}(\text{Males})$	0.786	0.810	0.799
$cc_{\uparrow}(\text{Females})$	0.848	0.861	0.853
$cc_{\uparrow}(G)$	0.817	0.835	0.826
$cc_{\downarrow}(\text{Males})$	0.456	0.486	0.470
$cc_{\downarrow}(\text{Females})$	0.226	0.255	0.241
$cc_{\downarrow}(G)$	0.321	0.370	0.337

Table 1 Generated Social Network Graph Statistics

Quantitative analysis of DynSNIC’s infection dynamic capabilities, in conjunction with health policies and interventions strategies, is provided in the following case study. 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 (Goldie, Kohli & Grima 2004). Upon acquisition of the HPV virus, the host could be asymptomatic for many years, clear the infection, or cervical dysplasia could develop. HPV prevalence is an integral component of cervical cancer’s etiology; although, DynSNIC’s vertex finite state machine is also capable of representing additional states beyond infection status, such as temporal pathogen dynamics (carcinogenesis). Presently, each vertex state machine in DynSNIC label HPV’s presence, susceptibility, or immunity (vaccination, other intervention or through cleared infection) in the host. We evaluate the impact of several disparate intervention strategies on HPV prevalence in the population. The simulator’s parameter space is gathered from (Corley & Mikler 2005). The demographic feature vectors are arbitrarily defined with five features, each feature with a integer value between 0 and 4. The discrete values are drawn from a uniform distribution. The use of a uniform distribution in the demographic features translates to a pseudo-”random” mixing due to the homogeneous nature of the population demographic strata composition. 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 (Corley & Mikler 2005).

Figure 4 HPV infectious population per gender / three intervention solutions. 10,000 pop size, 10 Monte-Carlo, 30 realizations each run

Population-level impact from three intervention strategies is evaluated; these include no intervention, vaccinating^a only males, and vaccinating only females. An intervention targeting both males and females would be economically cost-prohibitive and not included in our evaluations. Each Monte-Carlo simulation is loaded with the parameter space described earlier, population size of 10,000 ($|G_m| = |G_f|$), and executed for 30 discrete realizations (years). The impact of each intervention setting is averaged from ten Monte-Carlo simulations and the results are shown in Fig.4. 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 in female population) at the height of the epidemic when vaccinating females at 80% coverage and 75% efficacy. 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 with similar intervention solutions. 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 (Hughes, Garnett & Koutsky 2002, Sanders & Taira 2003). Endemic prevalence does not occur with our simulator; however, our results clearly show a reduction in prevalence within

^aIntervention coverage is 80%.

the RRP range of established models.

6 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 limited 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. With limited funding and resources available to help prevent infectious disease, public health professionals need tools to help them to make decisions regarding where the most effective measures would be taken. 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.

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, solely cosine similarity and an aggregate of the previous two as the contact likelihood. The next phase of development will assign social demographic feature distributions other than uniform, such as Gaussian or Poisson 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. One future case study is to evaluate demographic disparity in HIV/AIDS prevalence in the population and the effect of targeted public health information programs. This setting will incorporate behavioral data from national surveys; such as, the National Health and Social Life Survey (NHSL) the Center for Disease Control and Intervention's Youth Risk Behavior Surveillance Survey (YRBSS) and integrate concepts from information theory to study diffusion of information and the demographic-level consequences of that information, in the population (Laumann 1994, Centers for Disease Control and Prevention 2006).

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 $O(E \log V)$. 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.

7 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 and Notes

- Aral, S., Hughes, J., Stoner, B., Whittington, W., Handsfield, H., Anderson, R. & Holmes, K. (1999), ‘Sexual mixing patterns in the spread of gonococcal and chlamydial infections’, *American Journal of Public Health* **89**(6), 825–833.
- Barabási, A. & Albert, R. (1999), ‘Emergence of scaling in random networks’, *Science* **286**, 509–512.
- Barabási, A. & Albert, R. (2002), ‘Statistical mechanics of complex networks’, *Reviews of Modern Physics* **74**, 47–97.
- Berger-Wolf, T. & Saia, J. (2006), A framework for analysis of dynamic social networks, in ‘Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, Philadelphia, PA.
- Carley, K., Altman, N., Casman, E., Fridsma, D., Yahja, A., Chen, L., Kaminsky, B. & Nave, D. (2006), ‘Biowar: Scalable agent-based model of bioattacks’, *IEEE Trans. on Systems, Man, and Cybernetics* **36**, 252–265.
- Centers for Disease Control and Prevention (2006), Youth risk behavior surveillance - u.s. 2005, in ‘MMWR : Morbidity and Mortality Weekly Report’, Vol. 55 of *SS-5*, Center for Disease Control and Prevention, pp. 1–108.
- Chen, J., Kodagoda, D., Lawrence, M. & Kerndt, P. (2002), ‘Rapid public health interventions in response to an outbreak of syphilis in los angeles’, *Sexually Transmitted Diseases* **29**(5), 277–284.
- Corley, C. & Mikler, A. (2005), 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.
- Eames, K. & Keeling, M. (2002), ‘Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases’, *Proc Natl Acad Science* **99**(20), 13330–13335.

- Eng, T. & Butler, W. (1996), *The hidden epidemic*, National academy press.
- Erdos, P. & Renyi, A. (1959), ‘On random graphs’, *Publicationes Mathematicae* **6**, 290.
- Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z. & Wang, N. (2004), ‘Modeling disease outbreaks in realistic urban social networks’, *Nature* **429**, 180–184.
- Goldie, S., Kohli, M. & Grima, D. (2004), ‘Projected clinical benefits and cost-effectiveness of a humanpapillomavirus 16/18 vaccine’, *National Cancer Institute* **96**(8), 604–615.
- Gunn, R., Harper, B., Borntrager, D., Gonzales, P. & St.Louis, M. (2000), ‘Implementing a syphilis eliminations and importation contraol strategy in a low-incidence urban area: San diego county, ca 1997-1998’, *Am J Public Health* **90**, 1540–1544.
- Hamming, R. (1950), ‘Error detecting and error correcting codes’, *Bell System Technical Journal* **26**(2), 147–160.
- Hughes, J., Garnett, G. & Koutsky, L. (2002), ‘The theoretical population-level impact of a phrophylactic human papilloma virus vaccine’, *Epidemiology* **13**(6), 631–639.
- Jaccard, P. (1901), ‘-’, *Bulletin del la Socit Vaudoisedes Sciences Naturelles* **37**, 241–272.
- Kumar, R., Novak, J., Raghavan, P. & Tomkins, A. (2004), ‘Structure and evolution of blogspace’, *Communications of the ACM* **47**(12), 35–39.
- Latapy, M., Magnien, C. & Del Vecchio, N. (2006), ‘Basic notions for the analysis of large affiliation networks / bipartite graphs. arXiv.org:cond-mat/0611631’.
- Laumann, E. (1994), *The social organization of sexuality : sexual practices in the United States*, University of Chicago Press.
- Lewin, B. (1998), ‘Sex in sweden. on the sexual life in sweden 1996’, *Natl Inst Pub Health* .
- Liljeros, F., Edling, C., Amaral, L., Stanley, H. & Aberg, Y. (2001), ‘Human web of sexual contacts’, *Nature* **411**, 907–908.
- Newman, M. (2005), ‘Power laws, pareto distributions and zipf’s law’, *Contemporary Physics* **46**, 323–351.
- Newman, M. & Park, J. (2003), ‘Why social networks are different from other types of networks’, *Physical Review E* .
- Reiche, E. & et al. (2005), ‘Socio-demographic and epidemiological characteristics associated with human immunodeficiency virus type i(hiv-1) infection in hiv-1 exposed but uninfected individuals, and in hiv-1 infected patients from a southern brazilian population’, *Rev Inst Med trop San Paulo* **47**(5), 239–246.

- Rosenberg, D., Moseley, K., Kahn, R., Kissinger, P., Rice, J., Kendall, C., Coughlin, S. & Farley, T. (1999), 'Networks of persons with syphilis and at risk for syphilis in louisiana: Evidence of core transmitters', *Sexually Transmitted Diseases* **26**(2), 108–114.
- Salton, G. & McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill.
- Sanders, G. & Taira, A. (2003), 'Cost effectiveness of a potential vaccine for human papillomavirus', *Emerging Infectious Diseases* **9**(1), 37–48.
- Siek, J., Lee, L. & Lumsdaine, A. (2008), 'Boost c++ stl graph libraries'.
URL: <http://www.boost.org/libs/graph/doc/index.html>
- Ward, H., Ison, C., Day, S., Martin, I., Ghani, A., Garnett, G., Bell, G., Kinghorn, G. & Weber, J. (2000), 'A prospective social and molecular investigation of gonococcal transmission', *Lancet* **356**, 1812–1817.
- Watts, D. & Strogatz, S. (393), 'Collective dynamics of 'small-world' networks', *Nature* p. 440.
- Williams, L., Klausner, J., Whittington, W., Handsfield, H., Celum, C. & Holmes, K. (1999), 'Elimination and reintroduction of primary and secondary syphilis', *Am J Public Health* .
- Wylie, J. & Jolly, A. (2001), 'Patterns of chlamydia and gonorrhoea infection in sexual networks in manitoba, canada', *Sexually Transmitted Diseases* **28**(1), 14–24.
- Zipf, G. K. (1949), *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley.

