NDSSL Technical Report 09-066

July 08, 2009

Title: Structural and Relational Properties of Social Contact Networks with Applications to Public Health Informatics

Authors: Maleq Khan
V.S. Anil Kumar
Madhav Marathe
Zhao Zhao
Tridib Duta

Acknowledgements: We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory(NDSSL) for their suggestions and comments. This work has been partially supported NSF Nets Grant CNS- 0626964, NSF HSD Grant SES-0729441, CDC Center of Excellence in Public Health Informatics Grant 2506055-01, NIH-NIGMS MIDAS project 5 U01 GM070694-05, DTRA CNIMS Grant HDTRA1-07-C-0113 and NSF NETS CNS- 0831633.

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and State University

Structural and Relational Properties of Social Contact Networks with Applications to Public Health Informatics

Maleq Khan V.S. Anil Kumar Madhav Marathe Zhao Zhao Tridib Dutta
Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA
{maleq, akumar, mmarathe, zhaozhao, tridibd}@vbi.vt.edu

***Abstract*—It is now well established that social contact networks play a critical role in the progression of epidemics in urban regions. This paper analyzes the graphical and relational properties of large, realistic urban social contact networks. For purposes of exposition, we focus on Miami and Seattle; similar analysis was done for other large cities in the US. We compare these properties with a carefully chosen set of 3 other networks. Our results indicate that the structure of realistic social contact networks exhibit interesting differences amongst themselves as well as with other networks. As a byproduct, we develop a new algorithm for counting the number of different embeddings of a given (smaller) subgraph in a large sparse social network. This algorithm is an improvement over the previous algorithm for counting subgraphs efficiently in a large sparse graph.**

Next, we focus on a central question in Network Science – relationship between network structure and spread of diseases over these networks. We identify new structural measures that provide qualitatively different insights about epidemics; these insights can then be used for developing novel intervention strategies.

I. Introduction

The structure of social contact networks has been a topic of study in a number of areas, such as sociology [1], epidemiology [2] and data mining [3], [4]. Local and global network properties have been used in a number of applications, such as understanding the robustness of these networks [5], dynamics of spread of diffusion processes [6], formation of communities and clusters [3], and motifs and functional patterns in biological networks [7]. An important conclusion from the study of networks in a number of different areas is that they have very different structure from the classical Erdos-Renyi model, and they evolve over time. A number of models, e.g., [8], [5], have been proposed to capture their structure and evolution. However, most of these models have sought to match the degree distributions (and the clustering coefficient, to a lesser extent), and have not really considered other non-local properties.

The focus of this paper is: (i) *mining graphical and relational properties in large, realistic urban social contact networks*, and (ii) understanding how structural and relational properties (signatures) affect the spread of infectious diseases; we are in particular interested in identifying novel structural features that aid in policy planning as it pertains to pharmaceutical and non-pharmaceutical interventions. To the best of our knowledge, this is the first comprehensive study of its kind for social contact networks. Informally, social contact networks capture *proximity relationship* — vertices represent individuals and edges represent the fact two individuals were proximal for a certain period of time as they went about carrying their daily activities. Depending on specific diseases, the definition of proximity changes. Both the degree and clustering coefficient distributions can be viewed as specific instances of this approach - the degree distribution is a statistic of "maximal stars" and the clustering coefficient distribution is a statistic on the number of triangles in the graph. Our goal is to characterize the subgraph structure and understand the implications on global network properties, especially the disease dynamics. Subgraph mining is a very well studied area in data mining, in social, infrastructure and biological networks, though there are important differences with our work. One common topic is that of *frequent subgraph mining* [9], which involves determining the most frequent subgraph of a certain size. In biological networks, these have been used to identify *network motifs* [7] which have specific functional semantics. Distinguishing aspects of our work include (i) we focus on the enumeration of *all* subgraphs of certain structure and size, (ii) we study subgraphs in large scale social contact networks, which have not been explored at this scale (as discussed below), and (iii) we use demographic labels such as age and income level to aid in identifying properties. Most of the prior research on subgraph mining has been done on networks which model applications such as Recommender Systems [10], Blogs [4], Biological networks [7]. In these applications, there is a well defined network, which can be constructed, though with considerable difficulty (e.g., by crawling the web, in the case of the blog network, or through detailed biological experiments and bioinformatics analysis, in the case of biological networks). In contrast, social contact graphs, which model which individuals came into *physical* contact at some common location are extremely difficult to construct explicitly, as pointed out by a number of researchers [11]. Here we look at the patterns of social group formation in a large-scale synthetic social contact networks constructed by [12].

Our Results. We examine the structure of subgraphs in social contact graphs, as well as relational queries involving demo-

graphic labels. The different subgraphs we examine include *cliques*, *cycles* and *paths*, with different labeled attributes. We compare the social contact graphs with three different well studied graphs and models: the MySpace graph [13], the Internet autonomous system graph (AS) [14] and the preferential attachment (PA) model [8], and we find that these form a wide spectrum with social contact graphs and infrastructure graphs (i.e., AS and PA) on opposite extremes, with the MySpace graph in between. Our main results are summarized below. Some of the notation and terms used below are explained in Section IV.

1. We find that the counts of small subgraphs (e.g., cliques and cycles) in the synthetic social contact networks are much higher than in random graphs with the same degree distribution, and help in distinguishing them. However, these counts are difficult to compare across dissimilar graphs with different number of nodes and edges, and generalizing the commonly studied notion of clustering coefficient (CC), we introduce a new measure that we call the *k-clique coefficient (KCC)*; therefore, CC corresponds to $k = 3$. We find that as k increases, this measure is able to refine the characterization of social contact graphs - for instance, 3CC is able to separate social contact graphs from pure infrastructure networks (AS and PA), but the MySpace graph has a similar distribution for $k = 3$. In contrast, 4CC is able to separate social contact graphs from the MySpace graph as well. This suggests that enumerations of increasingly large subgraphs is able to improve the characterization of social contact graphs.

2. A unique feature of social contact graphs is the rich labeled structure and we find that labels have a significant effect on subgraph counts. For instance, subgraphs involving smaller children seem to be much more common. The label graphs also lend themselves to natural relational queries, some simple cases of which we translate to subgraph queries. We study simple queries, such as the average, minimum, maximum and gap between the ages of nodes in small cliques, and find that the age gap is low among such cliques. This helps understand the well known fact about increased transmission among children in the case of many diseases. We then explore to what extent subgraph counts can help in determining epidemic properties, such as vulnerability (which is, informally, the probability that a node would get infected - see Section IV for definitions), but we find no significant correlation between the vulnerability and the local subgraph counts, suggesting that fundamental epidemic properties such as vulnerability are inherently global. However, very simple relational queries on the dendograms, which are the (random) subgraphs formed by the edges on which the disease spreads, suggests very interesting observations, which can aid public health policy planning. The graphs we study also have edge labels, in the form of time of contact, and we examine the temporal properties of these graphs.

3. We develop a more efficient implementation of the subgraph enumeration algorithm by exploiting the labeled structure of our graphs, and using a careful ordering. We then use biased sampling to further improve the time complexity for approximate counting, based on the correlations between the subgraph counts and node degree.

In summary, the novel aspects of our work are the scale of the graphs we study (with millions of nodes), and the (node and edge) labeled queries, which help in public health informatics.

II. Related Work

Due to its wide applications, understanding social networks is important and caught attention from researchers in various areas such as sociologists [1], epidemiologists [2], and computer scientists [3], [15], [16], [6], [17], [10], [4].

A framework for identifying communities in dynamic (time-evolving) a social network is presented in [3]. A community is a group of closely interacting individuals where each individual interacts with the others in the group in a regular basis. An individual interacting with a group of people temporarily is not a part of the community – to be a part of the community, one must interact with the group in somewhat regular basis. They showed that identifying such communities is an NP-hard problem and presented heuristics based algorithm. It is not clear how well their algorithm scales to a large social networks, as their experiments was done on graphs with only hundred individuals.

Leskovec et al. [4] studied blog networks with the objectives of answering questions like: a) How do blogs cite and influence each other? b) How do such links evolve? c) How does popularity of the blogs drop? By crawling the blogs in the internet, they formed a blog network by adding an edge from blog post A to post B if A has a hyperlink to B; A and B can be two post in the same blog or in two different blogs. With the objective of understanding the underlying social network, they analyzed the directed graph formed by the aggregated links from blog posts to discover the patter of information flow in blogspace.

In [18], the authors presented a an algorithm to to find subgraphs of a graph matching a given query pattern. Their algorithms returns approximate matches where an edge can with a smaller path connecting the two nodes. The quality of the matches was evaluated using a goodness function. Their main contribution was the algorithm to find approximate matchings. They did not use it in analyzing the topological structure of a social networks.

Information cascades in a recommendation networks were analyzed in [11]. To understand the pattern of influence and the topological structure of the recommendation networks, they counts and enumerate the subgraphs that form the cascades. Their aim is to answer the questions like: What kinds of cascades arise frequently in real life? And how do they reflect properties of their underlying network environment?

III. Datasets

Our main focus in this paper is on social contact networks, in which a link represents physical contact between two people. Most of the research on social contact networks involves "explicit" networks, in which contacts exist explicitly - examples include blog networks [4], MySpace network [13] and recommendation system network [11]. Though, the data for these networks is difficult to get because of various reasons,

including privacy and security concerns, conceptually these networks are well defined. A study of epidemics (e.g., flu, which spreads by physical contact) requires social contact networks, in which an edge represents an actual physical contact between two people at some location during the day. Such networks do not exist, and are hard to construct exactly because it is even harder to track individuals' contacts over a day on a reasonable scale. In this paper, we study a specific class of synthetic social contact graphs, which have been built to address such difficulties. These networks, built using an urban transportation modeling and simulation environment called TRANSIMS [12], [19], [20], are statistically similar to real contact graphs, and have been calibrated by various indirect measurements. We briefly describe their construction and properties here; see [12], [19], [20] for more details. For comparison, we also study the corresponding properties in the following graphs: MySpace network [13], Internet Autonomous Level graph [14] and the preferential attachment model [5], [8].

TABLE I
NUMBER OF NODES AND AVERAGE DEGREES IN THE DATASETS

Dataset	Nodes	Avg. deg	Max deg
Miami	2,092,147	50.38	425
Seattle	3,207,037	55.35	456
MySpace	99,770	133.36	58,772
Internet	22963	4.22	2390
Pref. Att.(PA)	100,000	51.87	8299

TABLE II
NODE LABELS BASED ON AGE OR INCOME

Age Groups		Income Groups	
Age	Label	Income	Label
0-19	0	0-25K	0
20-39	1	25-50K	1
40-59	2	50-75K	2
60-79	3	75-100K	3
80-99	4	100-125K	4
99+	5	125-150K	5
		150-175K	6
		175-200K	7
		200K+	8

A. The TRANSIMS methodology for synthetic social contact networks

TRANSIMS [12], [19], [20] is a general framework for constructing synthetic populations and spatio-temporal models for urban traffic. We describe the model briefly below and refer the reader to [12], [19], [20] for further details due to lack of space. Step 1 creates a synthetic urban populations and is accomplished by integrating a variety of databases from commercial and public sources into a common architecture for data exchange that preserves the confidentiality of the original data sets, yet produces realistic attributes and demographics for the synthetic individuals. Each synthetic individual is placed in a household with other synthetic people and each household is located geographically in such a way that a census of our synthetic population yields results that are statistically indistinguishable from the original census data, if they are both aggregated to the block group level. In Step 2, a set of activity templates for households are determined, based on several thousand responses to an activity or time-use survey. These activity templates include the sort of activities each household member performs and the time of day they are performed. Thus for a city - demographic information for each person and location, and a minute-by-minute schedule of each person's activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques. This yields a *dynamic social contact network* represented by a (vertex and edge) labeled bipartite graph G_{PL}, where P is the set of people and L is the set of locations. If a person $p \in P$ visits a location $\ell \in L$, there is an edge $(p, \ell, label) \in E(G_{PL})$ between them, where $label$ is a record of the type of activity of the visit and its start and end points. It is *impossible* to build such a network by simply collecting field data; the use of generative models to build such networks is a unique feature of this work.

A substantial effort has been spent on calibration and validation of our relational networks; see (Barrett et al. 2001; Chowell et al. 2003; Eubank et al. 2004; Barrett et al. 2007) for details. First, the design of the system is based on a formal theory of simulation called Sequential Dynamical Systems (Eubank et al. 2005;Barrett et al. 2003; Barrett et al. 2007). Various microscopic and macroscopic quantities produced by TRANSIMS have been validated in the city of Portland, including (i) traffic invariants such as flow density patterns and jam wave propagation; (ii) macroscopic quantities, such as activities and population densities in the entire city, number of people occupying various locations in a time varying fashion, time varying traffic density split by trip purpose and various modal choices over highways and other major roads, turn counts, number of trips going between zones in a city, etc. Results on population mobility and social contact network construction were presented and reviewed annually (Barrett et al. 2001). The results were also reviewed in the context of epidemic modeling as a part of a letter report by the National Academies and published in (Halloran et al. 2008).

B. Preliminary properties of our datasets

We describe some basic structural and labeled properties of the graphs we consider. We use synthetic contact graphs for two different cities - Miami and Seattle, constructed using the TRANSIMS methodology, and compare them with the MySpace, Internet router graph and an instance from the preferential attachment model. Table III-B shows the number of nodes (people), average number of contacts (degree) and maximum number of contacts (maximum degree) for these contact networks, and Figure 1 shows the degree distributions. The demographic information about the Miami and Seattle contact graphs is shown in Table III-B. Figure 2 and 3 show the distribution of people based on age and income for these two cities.

IV. BACKGROUND: NOTATION AND PRELIMINARIES

Let $G = (V, E)$ be a social contact graph. We are given a set L of labels, with label $\ell(v) \in L$ for each node $v \in V$.

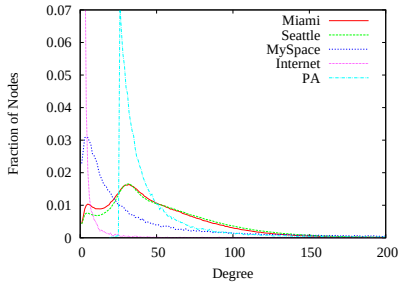


Fig. 1. Degree distribution of the contact networks

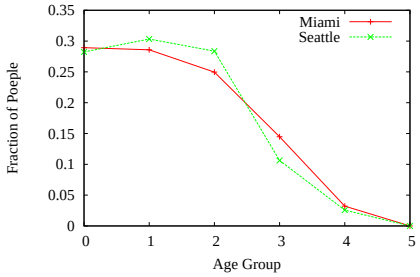


Fig. 2. Distribution of people based on age groups.

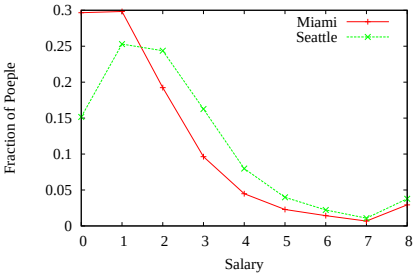


Fig. 3. Distribution of people based income groups.

The labels could denote demographic information, such as age, income, etc. In general, we can also have edge labels, though we only focus on node labels in this paper. Two labeled graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic* if there is a bijection $f : V_1 \to V_2$ such that (i) for each $u \in V_1$, $\ell(u) = \ell(f(u))$, and (ii) $(u, v) \in E_1$ if and only if $(f(u), f(v)) \in E_2$. The *graph isomorphism* problem is to find whether G_1 is isomorphic to G_2. The *subgraph isomorphism problem* is to find whether G_1 is isomorphic to some subgraph of G_2. In this paper, we consider the following counting version of the subgraph isomorphism problem.

Given a large graph $G = (V, E)$ and a relatively smaller graph $G_s = (V_s, E_s)$, count the number of subgraphs of G that are isomorphic to G_s. We describe the algorithm below.

We will study properties related to the spread of epidemics on networks. We focus on the SIR model of epidemics [5], [8], in which an *Infected* node u (in state I) spreads the disease to each *Susceptible* node v (in state S) independently, with some probability $p(u, v)$, and then *Recovers* (state R). Understanding the dynamics of disease spread, and methods to control it are important public health issues. A *dendrogram* is a labeled graph used to represent the stochastic output of an epidemic process on a graph. Formally, it is a tuple $R = ((V_0, \ldots, V_T, U), E_I)$, where (i) $V_0, \ldots, V_T, U$ is a partition of the set V of nodes, where V_t denotes the set of nodes that got infected at time t, and U is the set of nodes that never got infected, (ii) T denotes the last time any node got infected, and (ii) The set E_I denotes the set of edges on which the infection spread. Each dendrogram R for the graph G and a specific disease model appears with probability $p(R)$, which can be explicitly computed. For a dendrogram R, we let $t_R(v)$ denote the time at which node v got infected; note that $t_R(v) = \infty$ if $v \in U$, i.e., if node v never got infected in this run. A fundamental quantity related to disease dynamics is *vulnerability* $f(v)$ of a node v, which is defined as the probability that the node v becomes infected, if the disease starts at a random initial node [21]. Clearly, the vulnerability is a function of the disease model, especially the transmission probability, and the global graph structure.

Overlapping and Non-overlapping Subgraphs. Two subgraphs G_{s1} and G_{s2} of G are non-overlapping if they do not share any node. Two subgraphs overlaps if they have at least one node in common. We count all subgraphs and non-overlapping subgraphs. In all subgraphs, any pair of subgraphs may (or may not) overlap each other. To count non-overlapping subgraphs, we mark the nodes in subgraphs already found so that a marked node cannot be used again to match another subgraph.

Random Graphs with the Same Degree Distribution. We generate random graphs with the same degree distribution from a given social contact network by shuffling (aka swapping) the edges randomly: pick two edges (u_1, v_1) and (u_2, v_2) uniformly at random; then remove these two edges and add the edges (u_1, v_2) and (u_2, v_1) to the graph; repeat this process until all of the edges in the graph are shuffled. Thus, at the end, we get a random graph with the same degree distribution. After a certain number of iterations, it is easy to see that the Markov chain produces independent samples of the original graph with a given degree distribution. Section V-B discusses the resulting structure of these networks. Using this *graph null model*, we show that structural and relational properties of these realistic social contact networks cannot be expected by chance. In contrast the networks constructed in [22] are highly structured and have significantly different structural characteristics.

Clique Coefficient. We generalize the definition of clustering coefficient to a higher order measure of connectivity of neighbors of a node. Let $\text{NE}(v)$ be the number of edges among neighbors of node v and $d(v)$ be the degree of v. Clustering coefficient (CC) is defined as $\text{CC} = \frac{\text{NE}(v)}{\binom{d(v)}{2}}$, where $\binom{d(v)}{2}$ is the number of maximum possible edges among the neighbors of v. Notice that $\text{NE}(v)$ is also the number 3-cliques containing v and $\binom{d(v)}{2}$ is the maximum number of such cliques. In a similar fashion, we define k-clique coefficient (KCC) as $\text{KCC} = \frac{\text{number of k-cliques containing v}}{\binom{d(v)}{k-1}}$.

V. Subgraph Structure in Social Contact Graphs

We now describe our main results on the subgraph structure in social contact graphs. The main questions we focus on are: (i) What are the main characteristics of the counts of various kinds of subgraphs in social contact graphs, and how do they differ from other graphs, including random graphs. In particular, what are the characteristic differences that are not revealed through more conventional structural analysis, using properties such as the degree and clustering coefficient distributions; (ii) What is the effect of labels on subgraph

counts; (iii) how are subgraph counts related to properties of disease dynamics, such as node vulnerability. For most of these questions, we use only the Miami graph, since it seems to have very similar properties as the Seattle graph, as suggested by our initial analysis.

A. Clique Coefficient

Figure 5 and Figure 4 show the distributions of clustering coefficients (which is 3-clique coefficient (3CC) by our definition) and 4-clique coefficients (4CC), respectively. The distribution of clustering coefficients for the social contact networks Miami and Seattle is somewhat different than the other networks. However, the patterns of the distributions of 4-clique coefficients distinguish social contact networks Miami and Seattle much more clearly from the other networks. Thus, the higher order measure 4CC can be more useful over the traditional clustering co-efficient in characterizing some given types of networks; at least this is the case in the networks used in this paper.

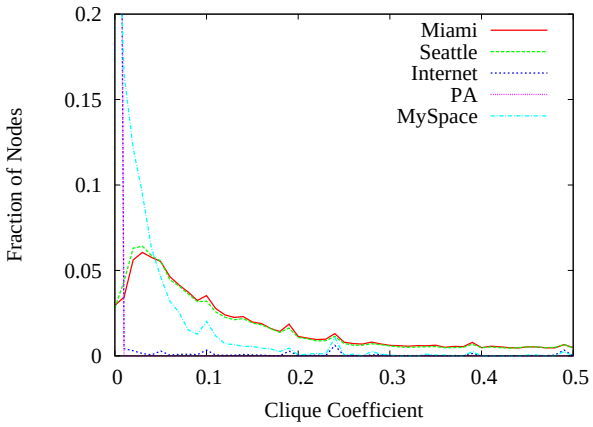


Fig. 4. Distribution of 4-clique coefficients.

In Internet and PA graph, for almost all of the nodes, 4CC is zero; thus we see a vertical line (in Figure 4) at zero for this two networks. In MySpace graph, although a very large of nodes have very small (zero or close to zero) 4CC, there are a significant number of nodes with larger 4CC. From this observation, we would like to conclude that with respect to clique coefficient, MySpace (an Internet-based social network) is showing an intermediate behavior in between infrastructure network Internet and real-life social contact networks Miami and Seattle.

Figure 6 and Figure 7 show average 3CC and 4CC, respectively, vs. degrees of the nodes: x-axis represent degree of the nodes and y-axis is the average 3CC and 4CC of all nodes having a given degree. For Miami and Seattle graph, both 3CC and 4CC is increasing first (up to degree 30) then decreasing sharply with degree. But for Internet and MySpace graph, both 3CC and 4CC are not changing that much with degree. Further, again, the contrast of Miami and Seattle to the other networks is much clearer with 4CC than 3CC.

B. Counts of labeled subgraphs and effect of random shuffles

First we show the number of non-overlapping subgraphs. A template subgraph of a age group means that all nodes in the subgraph are of that age group. Figure 8 shows the number of non-overlapping subgraphs in Miami graph for age group 0. As we are counting non-overlapping subgraphs, the number of subgraphs are decreasing with the size (number of nodes) of the subgraph. Obviously, the number of cliques is less than the number of stars, since a star is embedded in a clique of the same size, but not vice versa. However, the counts for all types of subgraphs (star, chain, cycle, and clique) are very close. We observe similar pattern for Age Group 1, 2, and 3. But, we found that the number of cliques is significantly smaller than the others for Age Group 4 (not shown in the Figure).

Figure 9 shows the counts of subgraphs for Age Group 0 after randomly shuffling all of the edges. After shuffling 100% edges, we get a random graph with the same degree distribution. In this random graph, the counts for stars and chains decreases by a very small amount. However, the number of cliques goes down to almost zero. The number of cycles shows an interesting behavior. After shuffling, the number of smaller cycles reduces significantly whereas the number of larger cycles goes down by smaller amount. Figure 10 shows the number of cycles for Age Group 0 with varying number shuffled edges. As more edges are shuffled, the number of smaller cycles are going down significantly.

Thus, the numbers of cliques and cycles clearly distinguish the social contact networks from this random graph even when the degree distribution remains same.

Figure 11 shows the number of overlapping subgraphs for Age group 0 and 3 in Miami network. The number of subgraphs is growing exponentially with the number of nodes in the template subgraph. Note that y-axis is in log scale. For Age Group 0 (unlike other age groups), the number of cycles is increasing at a faster rate than the number of stars. For a younger age group, the number of cliques is larger than that of an older age group as shown in Figure 12. A special interesting case is that, even though the number of people in Group 0 and 1 are almost equal (See Figure 2), the number of cliques for Age Group 0 is significantly larger (more than 10 times, take into consideration that y-axis in log scale) than that of Age Group 1.

Figure 13 shows the number of overlapping subgraphs with 100% edges shuffled. After shuffling, all of the subgraph counts become smaller. However, the number of cliques goes significantly down. In fact, the number of cliques of size larger than 3 is almost zero.

C. Correlations with degree and labels

We now examine how local node properties (e.g., the degree and labels, such as age and household income) affect the counts of subgraphs containing a specific node. Figure 14, 15, and 16 show distribution of average number of subgraphs of size 5 with respect to degree, age, and income, respectively. Let $count(v, s)$ be the number of overlapping subgraphs, isomorphic to s, that node v belongs to. In Figure 14, we plot average of the counts $count(v, s)$ of all vertices having a

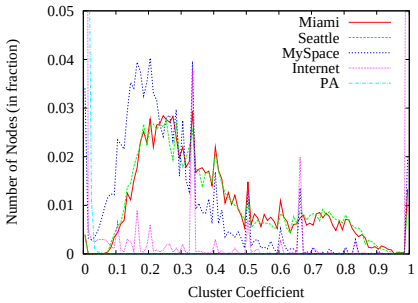


Fig. 5. Distribution of clustering coefficients.

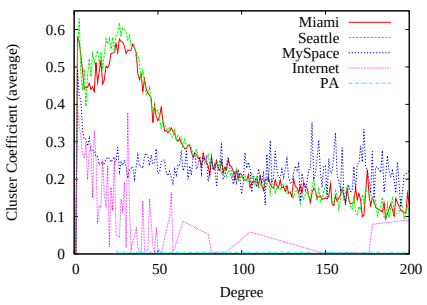


Fig. 6. Average clustering coefficient with degree.

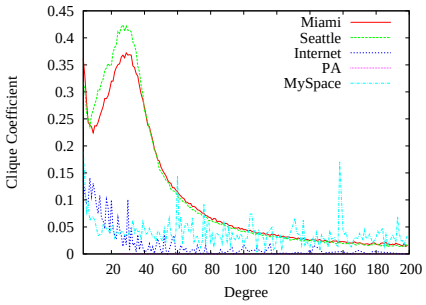


Fig. 7. Average 4-clique coefficient with degree.

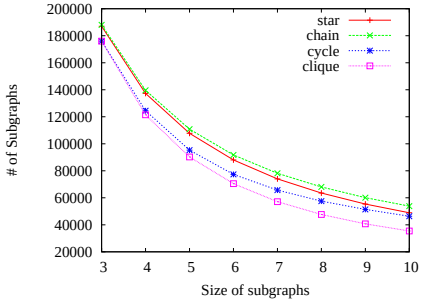


Fig. 8. Number of non-overlapping subgraphs in Miami graph for Age Group 0 (0-19).

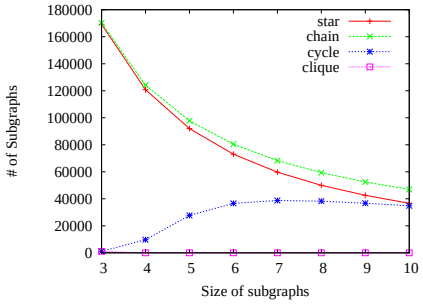


Fig. 9. Number of non-overlapping subgraphs in Miami graph for Age Group 0 (80-99) with 100% edges shuffled.

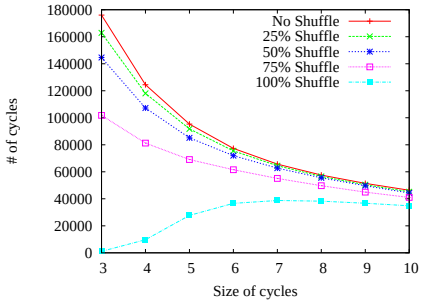


Fig. 10. Number of non-overlapping cycles in Miami graph for age group 0 with varying the number of shuffled edges.

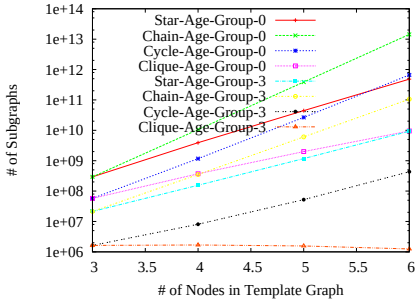


Fig. 11. Number of overlapping subgraphs in Miami graph for Age Group 0 & 3.

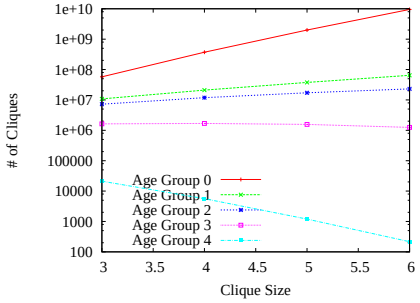


Fig. 12. Number of non-overlapping cliques in Miami graph for all age groups.

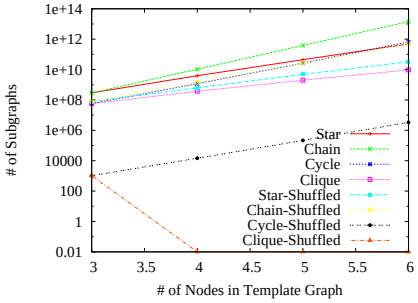


Fig. 13. Comparison of the subgraph counts in non-shuffled and shuffled graph (with 100% edges shuffled) for Age Group 0 in Miami graph.

particular degree. This plot shows that nodes of high degree tend to have much higher counts; note the contrast with the 4-clique coefficient (which is $count(v,s)/\binom{deg(v)}{4}$) in Figure 7, where high degree nodes have a low clique coefficient. An interesting observation is that the age (except youngsters, age < 20) and income do not seem to affect the counts at all. Figure 15 shows a youngster (with age < 20) is involved in more number of cliques (but in less number of stars and chains) than an older person. This fact is corroborated with an observation made in Figure 12 that the number of cliques for Age Group 0 is much larger than that of Age Group 1 even though the number of people in both groups are almost equal. Combining Figure 15 and Figure 12, we can conclude that the number of cliques for older age groups (say, Group 3 and 4) in Figure 12 is smaller is due to the only reason that there are less number of people in these older age groups.

D. Time Varying Dynamic Graphs

We also study dynamic social contact networks evolving over the time. We construct social contact networks at every hour beginning midnight for the next 24 hours. In a hourly network, there is an edge between two persons if they interact in that particular hour. Below we present the analysis of these 24 hourly contact networks for Miami city in an attempt to understand their dynamic evolution over the time. Figure 17 and 18 show how average degree and overlapping subgraph counts, respectively, varies with time of the day.

VI. Relational Mining and Public Health Applications

The labels in the synthetic social contact networks we study are a novel feature of these graphs and lend themselves to rich relational queries. We examine a class of simple

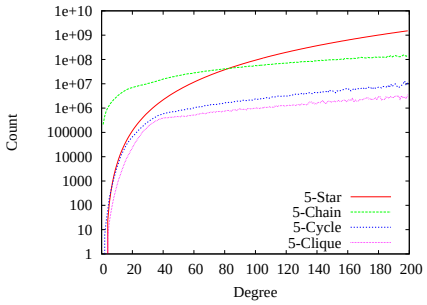


Fig. 14. Average number of overlapping subgraphs in Miami network w.r.t. degree.

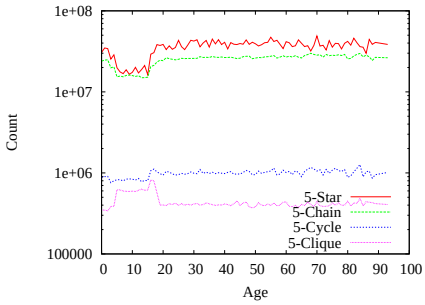


Fig. 15. Average number of overlapping subgraphs in Miami network w.r.t. age.

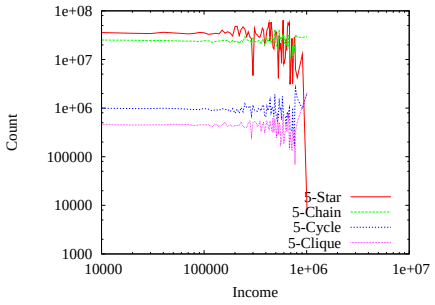


Fig. 16. Average number of overlapping subgraphs in Miami network w.r.t. income.

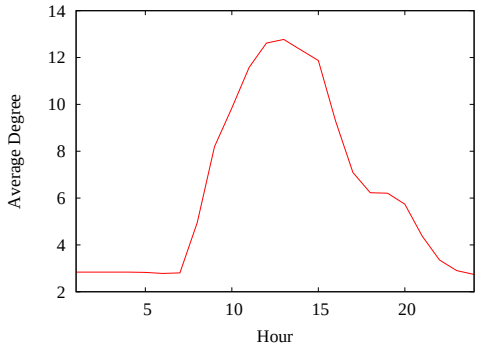


Fig. 17. Variation of average degree over time in Miami network.

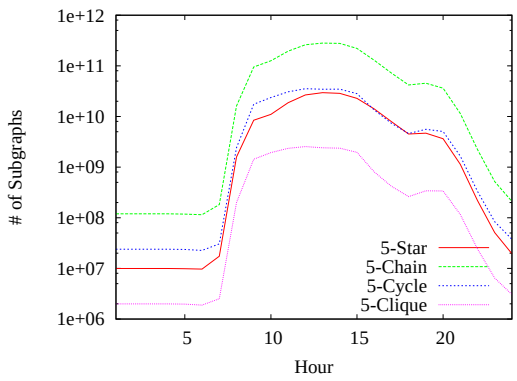
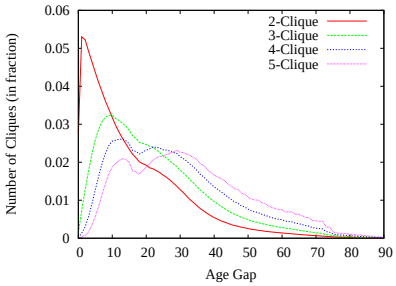


Fig. 18. Number of overlapping subgraphs with 5 vertices in Miami networks with time.

queries on labeled subgraphs with the goal of aiding public health policy planning. Let C be the set of vertices in a clique. Age(v) denotes the age of the person represented by vertex v. We consider the following simple functions defined on the set C: the minimum age, $min(C)$, the maximum age, $max(C)$, the average age, $avg(C)$ and the age gap, AgeGap$(C) = \max_{v \in C}$ Age$(v) - \min_{v \in C}$ Age(v); the results of such queries are shown in Figure 19. We note that Figure 19(b) was computed for a different synthetic graph for the state of Utah. The plots in Figure 19(a) suggest that more cliques seem to be formed by people of similar ages, and there seem to be generally more triangles among younger people (Figure 19(b)), since the average age is low. Disease transmission is known to happen more through children, which seems to be consistent with this observation of increased participation in cliques.



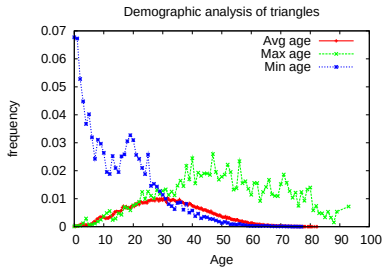


Fig. 19. (a) The AgeGap function on cliques of different sizes, (b) The average, min and max functions on triangles.

A natural question is to examine the extent to which disease properties of individuals can be inferred through subgraph counts. Such a relationship, if it exists, would be very valuable because the local neighborhood of a node can be determined much more accurately than the global network, and subgraph statistics (e.g., #s of stars and triangles) can be estimated easily. In order to explore this relationship, we plot vulnerability (recall the definition in Section IV) vs other local properties, such as subgraph counts and demographics. Unfortunately, we do not find any significant relationship between these quantities, as shown in Figures 20(a) and (b), and the vulnerability of an individual seems to be an inherently global quantity. However, we find that relational analysis of dendograms leads to interesting insights, even using very simple queries on 2-cliques, i.e., edges. In Figure 21, we compute the relative fractions of labeled edges in dendograms for the Portland contact graph; since each dendogram represents a random subgraph, the results in this figure are averaged over a number of dendograms. For each edge group, e.g., 10-19, we consider

all edges with one end point from this group, and we plot the frequency distribution of the ages of the other end point. Since these are dendograms, these edges represent actual stochastic disease transmission events. Note that these age groups are different from what have been considered in the rest of the paper (Section IV). We observe that except for the age group 10-19, for all other age groups, most of the transmissions are to adults with ages 20-40. For the group 10-19, most of the transmission are within the same group. This has very important implications for disease transmission and control policies for public health planners, because this suggests interventions could have different impact for different groups.

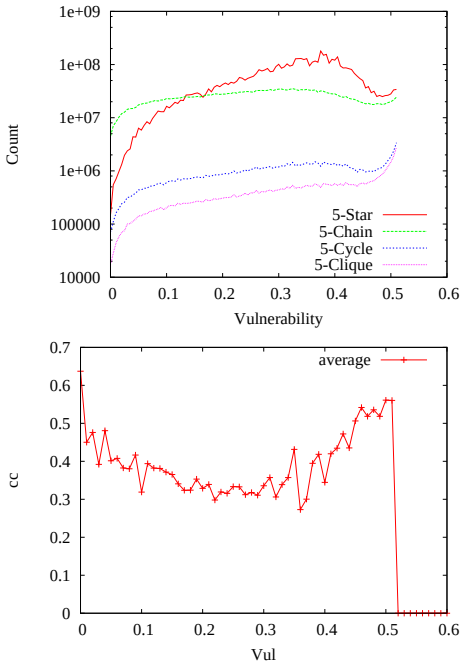


Fig. 20. Vulnerability vs (a) Subgraph count and (b) clustering coefficient for the Miami network.

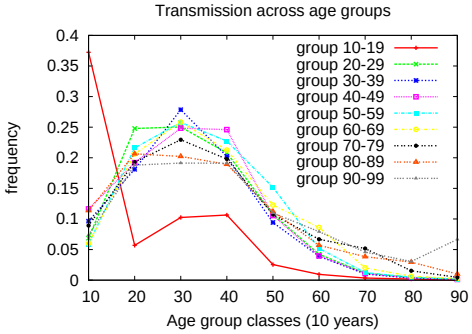


Fig. 21. Frequencies of labeled edges in dendograms.

VII. Subgraph Isomorphism Algorithm

There are many algorithms for subgraph isomorphism checking or finding a single matching subgraphs. However, counting all possible matching subgraphs is a more complex problem and have received less attention. Improving over the previous works, a more efficient algorithm for this problem was given in [23]. The basic strategy in our algorithm is same as that in [23]. We improve this algorithm further by reducing the number of candidate nodes to be matched at each recursive level of the algorithm.

The basic idea behind the algorithm is simple: fix a vertex u in the template subgraph G_s. If u can be matched with a vertex u in G, then recursively try to match the remaining nodes in $V_s - \{u\}$ with the nodes $V - \{u\}$ maintaining the structural integrity. Maintaining structural integrity means that for any two already-matched nodes $u_1, u_2 \in V_s$, if $(u_1, u_2) \in E_s$ then $(f(u_1), f(u_2)) \in E$.

Let $M_s \subset V_s$ be the set of nodes in G_s that has already been matched. Let $N(v)$ be the set of neighbors of v, and u be a node in G_s such that $u \in V_s - M_s$ and u has at least one neighbor in M_s. Define $M = \{f(u') \text{ — } u' \in M_s \wedge (u, u') \in E_s\}$. Then matching u with any node in $C = \bigcap_{v \in M} N(v)$ maintains structural integrity. Set C is called the set of *candidates* to match with u. Thus when all of the nodes in G_s are matched, we found an embedding of G_s in G. Below we provide the algorithm for counting all possible embeddings of G_s in G. The details are given below using procedures CountSubgraph (Algorithm 1) and MatchSubgraph (2).

Structural integrity is maintained by the way we compute the set of candidates C (described above). If the vertices are associated with labels, it is also necessary to maintain *semantic integrity*; that is, a vertex u in G_s with label $L(u) = x$ can only be matched with a vertex v in G with the same label $L(v) = x$.

Algorithm 1 Count Subgraphs of G that are isomorphic to G_s

CountSubgraph(G_s, G)
1: Perform a breadth-first search (BFS) in G_s. Let $u_1, u_2, \ldots$ be the nodes in G_s in the order of their discovery in BFS.
2: **for** each node $v \in V(G)$ **do**
3: **if** $L(u_1) = L(v)$) **then**
4: MatchSubgraph(2)
5: **return** count

Algorithm 2 MatchSubgraph(i)

MatchSubgraph(i)
1: **if** $i = |V_s| + 1$ **then**
2: $count \leftarrow count + 1$
3: **return**
4: Find $M = \{f(u_k) \text{ — } k < i \wedge (u_i, u_k) \in E_s\}$.
5: Compute $C = \bigcap_{v \in M} N(v)$
6: **for** each $v \in C$ **do**
7: **if** MatchVertex(u_i, v) **then**
8: MatchSubgraph($i + 1$)
9: **return**

Notice that in the above algorithm, if the template graph G_s has x automorphisms, G_s will be matched with same subgraph G'_s of G in x different ways. Thus the count returned by

$CountSubgraph$ must be adjusted by dividing it by x. The number of automorphisms x can also be computed using this algorithm by calling $CountSubgraph(G_s, G_s)$, which returns the number of possible embeddings of G_s in itself.

A. Sampling-based Estimation of Overlapping Subgraph Counts

The number of overlapping subgraphs grows exponentially with the number of vertices in the template subgraphs. Even counting the subgraphs of size 6 or 7 can take months in the large contact graphs (with millions of vertices) we are working with. We devise a simple sampling based approximation scheme, which can bring the running time from months to few days, even though this approximation scheme can take exponential time. Our approximation scheme is as follow.

In the above algorithm for counting subgraphs (Algorithm 1), we try to match u_1 with each node $v \in V(G)$. Let $v_1, v_2, \ldots, v_n$ be the vertices in G. Let c_i be the number of subgraphs isomorphic to G_s when u_1 is matched with v_i. Then the total count of the subgraphs is $c = \frac{1}{A(G_S)} \sum_{i=1}^{n} c_i$, where $A(G_S)$ is the number of automorphisms of G_s.

To have an estimation of this total count, we match u_1 with some selected v_is compute c_is for the selected v_is only. Let vertex v_i be selected with probability p_i, and X_i be an indicator random variable where X_i is 1 if v_i is selected and 0 otherwise. Then the estimated count is $\tilde{c} = \frac{1}{A(G_S)} \sum_{i=1}^{n} \frac{c_i}{p_i} X_i = \frac{1}{A(G_S)} \sum_{i:X_i=1} \frac{c_i}{p_i}$. The above estimator is unbiased as $E[\tilde{c}] = c$, since $E[\tilde{c}] = \frac{1}{A(G_S)} \sum_{i=1}^{n} \frac{c_i}{p_i} \Pr\{X_i = 1\} = \frac{1}{A(G_S)} \sum_{i=1}^{n} \frac{c_i}{p_i} p_i = c$.

Figure 22 shows the error in the estimations of the number of 4-cycles in Miami graph using i) uniform probability, $p_i = \frac{m}{n}$, where m is the expected number of samples, ii) probability proportional to node degree d_i, $p_i = \frac{d_i m}{d}$, where $d = \sum d_i$. Error is defined as $\frac{|c-\tilde{c}|}{c}$. Although intuitively, it seems that degree-based sampling would produce smaller variance, hence smaller error, we observe that sampling with uniform probability gives smaller error than sampling with probability proportional to node degree. With uniform probability, the error stabilizes at $m \geq 10,000$ (approximately 0.5% samples), and with $m \geq 20,000$ (1% samples) the error is smaller than 1%.

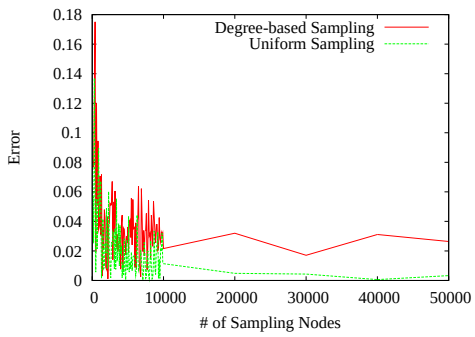


Fig. 22. Approximation error for the sampling algorithm.

VIII. Conclusions

We study the structure of subgraphs in large scale synthetic social contact graphs, as well as relational queries involving demographic labels. We define a new measure, the k-clique coefficient, which generalizes the clustering coefficient, and is able to characterize social contact graphs more tightly as k increases. We find that labels have a significant impact on the subgraph counts and simple relational queries which can be translated to subgraph queries lead to new insights for understanding the spread and control of epidemic processes on such networks.

References

[1] E. M. Rogers, *Diffusion of Innovations*, 5th ed. Simon & Shuster, Inc., 2003.

[2] L. A. Meyers, M. Newman, and B. Pourbohloul, "Predicting epidemics on directed contact networks," *Journal of Theoretical Biology*, vol. 240, p. 400418, 2006.

[3] C. Tantipathananandh, T. Y. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," in *KDD*, 2007, pp. 717–726.

[4] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2007.

[5] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[6] D. Kempe, J. M. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *ICALP*, 2005, pp. 1127–1138.

[7] N. Alon, P. Dao, I. Hajirasouliha, F. Hormozdiari, and S. C. Sahinalp, "Biomolecular network motif counting and discovery by color coding," *Bioinformatics*, vol. 24, no. 13, pp. i241–i249, July 2008.

[8] R. Pastor-Satorras and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2004.

[9] M. Kuramochi and G. Karypis, "Finding frequent patterns in a large sparse graph," *Data Mining and Knowledge Discovery*, 2005.

[10] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," in *Proceedings of the ACM Conference on Electronic Commerce*, 2006, pp. 228–237.

[11] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[12] C. L. Barrett, R. J. Beckman, K. P. Berkbigler, K. R. Bisset, B. W. Bush, K. Campbell, S. Eubank, K. M. Henson, J. M. Hurford, D. A. Kubicek, M. V. Marathe, P. R. Romero, J. P. Smith, L. L. Smith, P. L. Speckman, P. E. Stretz, G. L. Thayer, E. V. Eeckhout, and M. D. Williams, "TRANSIMS: transportation analysis simulation system," *Technical Report LA-UR-00-1725, Los Alamos National Laboratory*, 1997.

[13] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *World Wide Web (WWW) Conference*, 2007.

[14] "University of oregon route views project," http://routeviews.org/.

[15] J. E. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Natural communities in large linked networks," in *KDD*, 2003, pp. 541–546.

[16] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the web for emerging cyber-communities," *Computer Networks*, vol. 31, no. 11-16, pp. 1481–1493, 1999.

[17] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137–146.

[18] H. Tong, B. Gallagher, C. Faloutsos, and T. Eliassi-Rad, "Fast best-effort pattern matching in large attributed graphs," in *Proceedings of ACM SIGKDD*, August 2007.

[19] C. Barrett, S. Eubank, and J. Smith, "If smallpox strikes Portland?" *Scientific American*, vol. 292, 2005.

[20] S. Eubank, H. Guclu, V. S. A. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 180-184, 2004.

[21] C. Barrett, D. Beckman, M. Khan, V. A. Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis, "Generation and analysis of large synthetic social contact networks," in *Winter Simulation Conference*, 2009.

[22] N. Ferguson1, D. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. Burke, "Strategies for containing an emerging influenza pandemic in southeast asia," *Nature*, 2005.

[23] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–1372, October 2004.