Modelling structures of social groups

Shubin Mikhail

Department of Computer Science
PO Box 68, FI-00014 University of Helsinki, Finland
mikhail.shubin@helsinki.fi

Abstract. Models of social groups are mainly used in social sciences and epidemiology. There are many approaches to modelling social structures. This report gives an overview of such methods. Particular attention will be given to concept of social cohesion in graph model. This report shows the correlation between social cohesion and the k-connectivity of the model graph, describes a methods of learning different aspects of social groups (such as power or attachment to school) from to the properties of graphs. Advantages and disadvantages of the proposed method are discussed.

1 Introduction

Mathematical and computer modelling is an essential part of science today. This report is devoted to modeling of population and social structures.

Population models have a significant influence in current sociological research and theory. They are used in theory of economic sociology or stratification, they been used to describe social support, processes in health and health policy [13], family demography [14], and in the analysis of criminal networks[15]. In epidemiology such models can be used to model spread of infections, predict the peak and outbreaks of the epidemic, efficiency of vaccination[2], school closure[16], quarantine and other methods of epidemic control.

I will mainly report on article "Structural Cohesion and Embeddedness" by J. Moody and D. R. White [1], witch describe the correlation between social cohesion and the k-connectivity of the model graph. I will also use several examples from the articles, devoted to modelling of the swine flue epidemics.

In the second second section of report an overview of different methods of modelling population will be given; third section covers the use of graph in modelling of social groups; forth section deals with result of modelling; fifth section gives an overview of graph algorithms, utilized in social modelling; sixth section is a discussion.

2 Different methods of modelling population

Modelling of population is important problem. Many method for modelling population have been developed. In this section several approaches from mathematical epidemiology will be described.

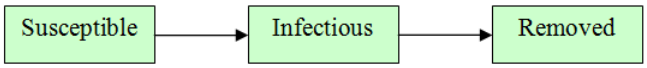

Fig. 1. SIR model

One of the approaches for modelling epidemics of inflectional disease is to use differential equations. One of the most simple and widely used model is SIR (Susceptible Infectious Removed) model. In this model population divided into 3 categories (figure 1) , with individuals moving from category 'susceptible' to 'infectious', and from 'infectious' to 'removed'. Mean number of people in each category can be described with the system of 3 differential equation.

$$\frac{dS}{dt} = -\alpha SI$$

$$\frac{dI}{dt} = \alpha SI - \beta I$$

$$\frac{dR}{dt} = \beta I$$

Where S - number of suspected, I - number of infectious, R - number of removed, α - infectivity and β - removal rate [1].

SIR model does not consider social structure at all, it uses only total number of individuals in the population. In this case population is homogeneous. For example, we can say that our population consists of S=1000 persons, and then admit infection with certain parameters to them.

Several variations can be introduced to SIR model to describe heterogeneity (differences) inside the population. We can divide population by age, sex, geographical position, susceptibility, contact rate, etc.

In metapopulation approach [3, 4] the world is divided into geographical regions defining a subpopulation network where connections among subpopulations represent the individual fluxes due to the transportation and mobility infrastructure. metapopulation model has two levels. The population layer is based on the high-resolution population database of the 'Gridded Population of the World' project of the SocioEconomic Data and Applications Centre (SEDAC) that estimates the population with a granularity given by a lattice of cells covering the whole planet at a resolution of 15x15 minutes of arc. The transportation mobility layer integrates air travel mobility obtained from the International Air Transport Association (IATA) and Official Airline Guide (OAG) databases that contain the list of worldwide airport pairs connected by direct flights and the

[1] Removed individual are either dead, or recovered individual with 100% immunity. In basic SIR model there is no difference between those two outcome of the illness, because they affect spread of the infection in the same way

number of available seats on any given connection. The combination of the population and mobility layers allows the subdivision of the world into georeferenced census areas defined with a Voronoi decomposition [2] procedure.

One interesting model with high level of complexity is presented in the article [2]. Whole describtion of the model is quoted here to illustrate assumption and data sources for this kind of research.

The population is divided into census tracts[3], and each tract is subdivided into communities of 500 – 3000 individuals. Each community is populated by randomly generated households of size 1 – 7 using the US-wide family size distribution from the 2000 Census. The household is the closest social mixing group, within which contacts between individuals occur most frequently and thus influenza is transmitted most often. The population is organized as a hierarchy of increasingly large but less intimate mixing groups, from the household cluster (sets of four socially close households), neighbourhoods (1/4 of a community), and the community. Including such groups creates a realistic contact network for disease transmission. At night, everyone can make contact with other individuals in their families, household clusters, home neighbourhoods, and home communities. In the daytime, individuals might interact with additional groups. During the day, most children attend school or a playgroup, where there is a relatively high probability of transmission. Preschool-age children usually belong to either a playgroup of four children or a neighbourhoods preschool, which typically has 14 students. Each community has mixing groups that represent two elementary schools, one middle school, and one high school, which typically have 79, 128, and 155 students, respectively.

Most working-age adults (about 72% of 19-64 year-olds) are employed. Employment rates are determined on a tract-by-tract basis using data from the US Census 2000. Employed individuals often work outside of their home communities. Each employed individual is assigned to work in a destination census tract based on commuting data taken from Part 3 of the Census Transportation Planning Package. Working individuals are assigned to communities and neighbourhoods within their destination tracts to simulate casual community contacts during the day, and a work group of about 20 people to represent their close contacts at the workplace. Unemployed individuals remain in their home communities and do not have close daytime contacts except with members of their households who are not employed or enrolled in school.

Individuals can engage in short-term, long-distance domestic travel to represent vacations and other trips. The traveller will stay at the destination for 0–11 nights, with 23.9% of trips lasting for a single day (and no nights), 50.2% including 1-3 nights away, 18.5% including 4-7 nights away, and 7.4% for 8-11 nights. A random member of this community is assigned to be the traveller's contact person, and at night the traveller will behave as if he/she belongs to the

[2] Given continuous space S ans a set of dots $d_1 \dots d_n$, Voronoi decomposition breaks space into n areas so that each dot d_i have "received it's own" area A_i (all dots in A_i have d_i as the closes dots among $d_1 \dots d_n$)[18]

[3] census tract is a geographic region defined for the purpose of taking a census

contact's household, household cluster, and neighbourhood. The traveller may withdraw to this household if ill.

Simulation epidemic on the model with has been used to study pathways of of epidemic, infection rate on different stages, efficiency of different types of vaccination and, school closure, ets [2]. Figure 2 illustrate prevalence of the infection in the simulation of Seattle with different types of epidemic control strategies.

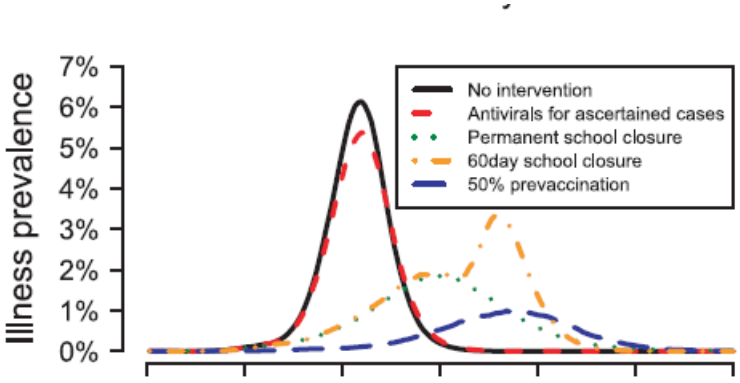

Fig. 2. Resultate of simulation of flu epidemic in population model [2]

3 Using graphs for modelling social groups

Graph models is a obvious method for description of structures with interconnections. It can be used to model population with complicated social structure.

Consider the graph model, where each node of the graph represents the actor and the edge represents relations between two actors [4]. Type of relation can differ to fit the purpose of the model.

We take several examples to illustrate the concept.

Consider group of five people. Figure 3.a shows a situation where all the actors(individuals) communicate indirectly. We can imagine this kind of relations forming among a collection of isolated individuals. Fig 3.b shows a group united around one strong charismatic leader (Each person might have ties to the leader, and be connected only through the leader to every other member of the group) and Fig 3.c shows a group where all actors communicate with each other.

Different sources can be used to build a model. Most studies use data from sociological articles, commercial databases or social networks.

[4] Actor is a unit of social relation. It can be individual, family, organization, ets., depending on the scale of the research. Graph consisting of different type of actors can be utilized. In most of the following examples we assume that actor is an individual.

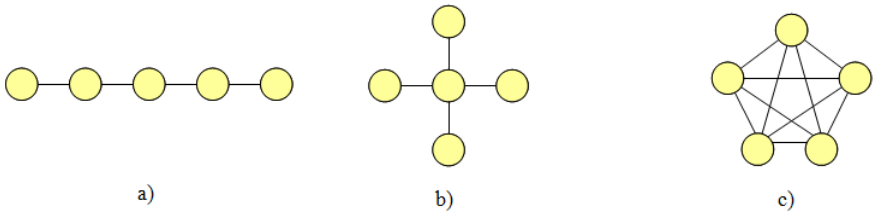

Fig. 3. Example of graph models of social groups

4 Analyzing graph model of the social structure

This section describes several properties of social group that we can learn by studying model of this group.

4.1 Social cohesion

Social cohesion can be preliminary defined as a field of forces that act on members to remain in the group [5]. In Social science, variety definition of cohesion are used. J. Moody and D. R. White have studied terminology and separated five most important features of definitions[1].

1. Cohesion is a property describing how a collection of actors is united.
2. It is a property of the group. Individuals may be embedded more or less strongly within a cohesive group, but the group has a unique level of cohesion.
3. This concept is continuous. Some groups are weakly cohesive (not held together well), while others are strongly cohesive.
4. Structural cohesion rests on observable social relations among actors.
5. The definition makes no reference to group size.

They pointed out disadvantage of all this definitions: they can not be used as an exact measurement of cohesion. The are too vague and give no approach to compute cohesion numerically.

J. Moody and D. R. White introduced their own definition of cohesion, witch unite all five main features and can be represented with exact number.

> **Definition 1:** *A group's structural cohesion is equal to the minimum number of actors who, if removed from the group, would disconnect the group*[1].

Consider the example from Figure 3. Removal of one connection from (a) or (b) cause the group to fall apart. In opposite, 4 connection should be removed from (c) to break it, thus we consider (c) to be more cohesive.

This property corresponds to k-connectedness.

Definition 2: *A graph G is said to be k-connected if there does not exist a set of k vertices whose removal disconnects the graph*[10].

Due to Menger's Theorem[5], we can alternatively define structural cohesive as number of independent relational paths among all pairs of members. The presence of multiple paths, passing through different actors, implies that if any single actor is removed, alternative links among members still maintain social solidarity.

4.2 Finding a subgroups inside a social group

Because formal specification for structural Cohesion have been created, we can link network structure to actor mechanisms to derive further theoretical consequences of structural cohesion. We can find core subgroups inside a social groups using such a parameter of a graph as nestedness.

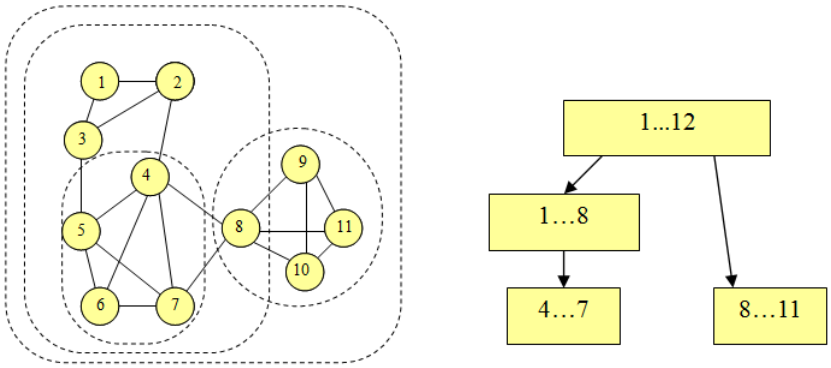

Fig. 4. Nestedness and it's hierarchical representation

Consider the example given in Figure 4. This network has a single component inclusive of all nodes. Embedded within this network are two biconnected components: nodes {1..8} and {8-11}, with node 8 involved in both. Within the first bicomponent, however, members {4-7} form a 3-component (a four-person clique) and all the members of second becomponent {8-11} form a 3-component. Thus, the group structure of this network contains a 3-level hierarchy, which is presented in Figure 4.

Learning nestedness of population can be very important in mathematical epidemiology. As example, sexual transmited deciases spread though hightly sexual active cores incide population.

[5] Let G be a finite undirected graph and x and y two nonadjacent vertices. Menger's Theorem states that the minimum number of vertices whose removal disconnects x and y is equal to the maximum number of pairwise vertex-independent paths from x to y. [17]

4.3 Informational and resource flow

Information and resources can flow through multiple paths, making control of resources within the group by a small ($\leq k$) number of people difficult. Although many potential implications likely follow in particular substantive areas, we focus below on three broad types of sociological questions: resource and risk flow, community and class formation, and power.

A focus on structural cohesion provides new insights into diffusion, augmenting current approaches that focus largely on network distance. The length of a path (number of edges) is often considered critical for the flow of goods through a network, as flow may degrade with relational distance. That is, the probability that a resource flows between two actors is equal to the product of transition probability of all edges along the path(s) connecting them. When multiplied over long distances, the efficacy of the information diminishes, even if the pairwise transmission probability is high. For example, the probability that a message will arrive intact over a six-step chain (This is the purported average acquaintance distance among all people in the United States [6]) when transmission probability of all edges is 0.9 will be 0.53. The fragility of long-distance communication rests on the fact that at any step in the communication chain, one person's failure to pass the information will disrupt the flow. For a structurally cohesive group, however, expected information degradation decreases with each additional independent path in the network. For example, the comparable probability of a six-step communication arriving given two independent paths is 0.78. In a high-connectivity network, even if many people stop transmission (effectively removing themselves from the network), alternate paths provide an opportunity for spread.

4.4 Community and Class Formation

Structural cohesion provides us with a useful tool for understanding processes related to the formation of social classes, ethnicity, and social institutions. Linking structurally cohesive subgroup membership to institutions that provide formal access to power suggests a new approach to the study of social stratification and the state. D. White et al. [9], for example, identify an informally organized 'invisible state' created by the intersections of structurally cohesive groups across multiple administrative levels. They show that those who share administrative offices during overlapping time spans build dense cliquelike social ties within a political nucleus while maintaining sparse locally tree-like ties with structurally cohesive groups (globally multiconnected) in the larger region and community. The locally dense and the globally sparse multiconnected ties act as different kinds of amplifiers for the feedback relations between larger cohesive groups and their government representatives. In his classic statement on the development of social capital, Coleman [8] argued that a closed-loop structure connecting adolescents - friends - parents increases effective normative regulation in a community. The key structural feature responsible for this increased ability is that biconnected components (loops) allow information to flow freely throughout the community, allowing normative ideas to be exchanged and reinforced.

4.5 Power

The substantive character of groups that are vulnerable to unilateral action differs significantly from that expected of groups with multiple independent connections. The group as a whole is vulnerable to the will and activities of those who can destroy the group by leaving. Moreover, actors that can disconnect the group are also actors that can control the flow of resources in the network. As has long been known from Network Exchange Theory, networks with structural features leading to control of resource flows generate power inequality [7].

4.6 Examples

Three examples demonstrate the empirical validity of a structural conception of social cohesion.

4.6.1 Cohesion among large American businesses It is possible to approach the question of business unity as a problem of structural cohesion. Because structural cohesion facilitates the flow of information and influence, coordinated action, and thus political activity, ought to be more similar among pairs of firms that are similarly embedded in a structurally cohesive group. Mizruchi [11] highlights the importance of financial institutions for unifying business activity. He identifies the number of indirect interlocks between two firms as the number of banks and insurance companies that have direct interlocks with both manufacturing firms in the dyad.

Data on large manufacturing firms can be used to identify the cohesive group structure based on indirect interlocks and relate this structure to similarities in political action.

The sample Mizruchi constructed consists of 57 of the largest manufacturing firms drawn from the twenty major manufacturing industries in the U.S. Census Bureau's Standard Industrial Classification Scheme. In addition to data on directorship structure, he collected data on industry, common stock-holding governmental regulations, and political activity. The question of interest is whether the structure of relations among firms affects the similarity of their behaviour. To explore whether firms that are similarly embedded also make similar political contributions, Mizruchi constructed a dyad-level political contribution similarity score as a function of the number of common campaign contributions. He modeled this pair-level similarity as a function of geographic proximity, industry, financial interdependence, government regulations, and interlock structure.

A cohesive blocking of this network reveals that most firms are involved in a strongly cohesive group, with 51 of the 57 firms members of the largest bicomponent. The nestedness structure consists of a single hierarchy that is 19 layers deep, and at the lowest level (at which no further minimum cuts can be made that would not isolate all nodes), 28 firms are members of a 14-connected component (the strongest k-component in the graph).

Studies shows that joint membership in more deeply nested subsets lead to greater similarity in political contributions [1]

4.6.2 Structural Cohesion in Adolescent Friendship Networks Add Health is a school-based study of adolescents in grades 7 through 12. A stratified nationally representative sample of all public and private high schools (defined as schools with an 11th grade) in the United States with a minimum enrolment of 30 students was drawn from the Quality Education Database (QED) in April 1994 [12]. Network data were collected by providing each student with a copy of the roster of all students for their school. Students identified up to 5 male and 5 female (10 total) friends from this roster. Data on more than 4,000 students have been taken from a dozen schools with between 200 and 500 students (mean = 349) provided a diverse collection of public (83 percent) and private schools from across the United States.

For each school, cohesive blocking procedure were employed to identify all connectivity sets for each school friendship network. At the first level, we have the entire graph, which is usually unconnected (because of the presence of a small number of isolates). Most of the students in every school are contained within the largest bicomponent, and often within the largest tricomponent. More of smaller and more tightly connected groups are identified. In these data, at high levels of connectivity ($k > 5$), subgroups do not overlap. This implies settings with multiple cores, differentially embedded in the overall school networks.

Nestedness within the community is reflected in a student's perception of his or her place in the school. The Add Health in-school survey asks students to report on how much they like their school, how close they feel to others in the school and how much they feel a part of the school.

Several correlation can be found. First, the number of contacts a person has (degree centrality) reflects their level of involvement in the network. Substantively, those people with many friends in school are more likely to feel an integrated part of the school. Second, having friends who are all friends with one another is an important feature of network involvement. As such, the density of one's personal (local) network is tested. Third, those people who are most central in the network should have a greater sense of school attachment. [1]

4.6.3 Shortest part between facebook users On-line social networks can serve as an data source for graph models, and can be object of the study themselfs. Researchers from University of Cambridge proved that in social network facebook.com length of maximal shortest part between users can be aproximated as only 3. [20]

5 An algorithm for finding nestedness and k-components

Combining algorithms from computer science we can identify cutsets in a network as follows:

input data: non weighted finite undirected graph
output: tree-like structure, describing nestedness

1. Identify the connectivity, k, of the input graph.

2. Identify all k-cutsets at the current level of connectivity.
3. Generate new graph components based on the removal of these cutsets (nodes in the cutset belong to both sides of the induced cut).
4. If the graph is neither complete nor trivial, return to 1; else end.

This procedure is repeated until all nested connectivity sets have been enumerated.

Testing for k-connectivity (Step 1) can be accomplished with a network maximum flow algorithm (look [19] for more information).

Two steps can be taken to reduce the computation time. First, there are linear time algorithms for identifying k-connected components for $k \leq 3$, and one can start searching with these algorithms, limiting the number of levels at which one has to run the full connectivity algorithms. Second, in many empirical networks the most common cutset occurs for singleton cuts. Because the procedure is nested, one can search for nodes with degree less than or equal to the connectivity of the parent graph (the graph from which the current graph was derived), remove them from the network, and thus apply the network flow search only after the singleton cuts have been removed

6 Discussion

We have given an overview of models of population, described a of graph model. We showed the correlation between social cohesion and the k-connectivity of the model graph, described a method of learning different aspects of social groups (such as power or attachment to school) from to the properties of graphs. This method have several advantages.

Method is easily scaled. But on the practical level many practical problems can be predicted. First, it is almost impossible to find any data for building a reliable large scale model. Second, large scale model will be computationally hard. This kind of problems are common for all large-scale researches.

Method provides the strict definition for structural cohesion and embeddedness, witch can be useful in many situations. neighbourhoods is based on graph theory, witch is well developed. Efficiency of the method is proved on several empirical examples. But should remember that in social sciences, vague definition can some times work better then strict one, and exact numerical representation can be redundant or even misleading. As example, we will not seriously consider the idea of defining beauty with mathematical formulas, or measuring the happiness.

Another disadvantage of the approach: it use only 1 type of interaction between individuals. Interaction can only exist, or not exist. Of course, this model does not cover variability in communication between people. Answer to the question 'how many friends you have' can vary greatly, depending on the definition of 'friendship'. 'Power' of interaction between individuals can be essential in most of practical problems. As example, to model the flue epidemic, we need take into consideration not the number of people in contact with infective person, but the rate of these contacts.

To solve this problem, model can be expanded by using weighted graph instead of nonweighted one. But computation difficulty will be greatly increased.

Although some of this modellings can be useful, we always should keep in mind that the reality is much more complicated then the model, and real life can not be reduced to numbers.

I expect that proposed model can work with the best efficiency within the following limitation.

1. Using of binary interaction should be theoretically substantiated.
2. Model preferably to be small sized.

Several counterexamples can be introduced. Army have a strong hierarchical control, so model of relation within army will be represented by the tree graph. Thus, k-connectivity of the model will be 1, and army should have very low structural cohesion due to definition. But in real life situation is completely different.

Several way for future research can be proposed.

1. Approach can be expanded to the case of weighted graph (variation of relation types)
2. To count social cohesion several other mathematical concepts, similar to k-connectivity, can be utilized. As example, graph compression[16] can be used to find strongly cohesive cores inside social groups.
3. Similar theory for studying highly hierarchical structures (such as army) can be developed.

7 References

1. James Moody, Douglas R. White: Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. American Sociological Review 68, 2003.
2. FluTE, a Publicly Available Stochastic Influenza Epidemic Simulation Model Dennis L. Chao, M. Elizabeth Halloran, Valerie J. Obenchain, Ira M. Longini, Jr. 1 Center for Statistics and Quantitative Infectious Diseases/Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 2 Department of Biostatistics, School of Public Health/University of Washington, Seattle, Washington, United States of America. 2009.
3. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose J Ramasco, Daniela Paolotti, Nicola Perra, Michele Tizzoni, Wouter Van den Broeck, Vittoria Colizza and Alessandro Vespignani. 2009.
4. Rvachev LA, Longini IM: A mathematical model for the global spread of influenza. Math Biosci 1985.
5. Festinger, Leon, Stanley Schachter, and Kurt Back. Social Pressures in Informal Groups: A Study of Human Factors in Housing. Stanford, CA: Stanford University Press. 1950.

6. Milgram, Stanley. The Small World Problem. Psychology Today 1969.
7. Willer, David. Network Exchange Theory. Westport, CT: Praeger. 1999.
8. Coleman, James S. Social Capital in the Creation of Human Capital. American Journal of Sociology 94 (Supp.) 1988.
9. White, Douglas R., Michael Schnegg, Lilyan A. Brudner, and Hugo Nutini. Conectividad Multiple, Fronteras e Integracin: Compadrazgo y Parentesco en Tlaxcala Rural (Multiconnectivity, Social Boundaries and Integration: Ritual Sponsorship and Kinship in Rural Tlaxcala). Pp. 4194 in Anlisis de Redes: Aplicaciones en Ciencias Sociales, edited by J.G. Mendieta and S. Schmidt. Mexico City, Mexico: Instituto de Investigaciones en Mathemticas Aplicadas y en Systemas de la Universidad Nacional Autnoma de Mxico. 2002
10. Harary, F. Graph Theory. Reading, MA: Addison-Wesley, 1994.
11. Mizruchi, Mark S. The American Corporate Network, Beverly Hills, CA: Sage. 1982
12. Bearman, Peter S., Jo Jones, and J. Richard Udry. Connections Count: The Add Health Design. 1996. Retrieved January 7, 2003 (http://www.cpc.unc.edu/projects/addhealth/design.html).
13. Healy, Kieran. Emergence of HIV in the US Blood Supply: Organizations, Obligation, and the Management of Uncertainty. Theory and Society 1999.
14. Astone, Nan Marie, Constance A. Nathanson, Robert Schoen, and Young J. Kim. Family Demography, Social Theory, and Investment in Social Capital. Population and Development Review 1999.
15. Baker, Wayne E. and Robert R. Faulkner. The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. American Sociological Review 1993
16. Navlakha, S., Rastogi, R., and Shrivastava, N.: Graph summarization with bounded error, SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ACM, 419432, 2008.
17. Aharoni, Ron and Berger, Eli, Menger's Theorem for infinite graphs. Inventiones Mathematicae 2009
18. Franz Aurenhammer, Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. ACM Computing Surveys, 1991.
19. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, NJ, 1993.
20. Joseph Bonneau, Jonathan Anderson, Ross Anderson, Frank Stajano, Eight Friends are Enough: Social Graph Approximation via Public Listings, f Second ACM Workshop on Social Network Systems, 2009.