Univerza *v Ljubljani* Fakulteta za *matematiko in fiziko*



Oddelek za fiziko

SEMINAR Computational Social Science

Author: Andrej Kranjec

Mentor: prof. dr. Rudolf Podgornik

Abstract

Computational social science is becoming an ever more interesting research field for social scientists, mathematical sociologists, computer scientists and even physicists. Network theory provides a simple and useful framework for studying complex social systems. Physicists have contributed fundamental work to complex network research with a unique perspective on large complex systems using concepts from statistical physics. Advances in computer technology have enabled the empirical study of large social networks, which have been previously limited to relatively small groups of people.

A theoretical basis for describing networks is presented with an emphasis on social network properties. Three concrete studies of a large social network providing empirical results for different aspects of human behavior are described in detail. The commonalities and differences in network behavior for each study offer an insight on individual level behavior, which is briefly discussed. An outlook of the potential of such empirical studies implementing new technologies to acquire large data sets for social networks is also briefly speculated.

Contents

Introduction	1
Basics of Network Theory	2
Characterization of Networks	2
Network Classification	3
Social Networks	5
Examples of Social Network Studies	6
Source Data and Network Ascertainment	6
Statistical Analysis	6
Results	7
The Spread of Happiness	7
The Spread of Loneliness	10
The Spread of Obesity	12
Discussion	. 14
Conclusion	. 15
References	.16

Introduction

Society is a large complex system of individuals, each with his own complex personality and behavioral patterns. This personal complexity has been the main interest of social scientists in the past. By recognizing general properties in different personalities, they constructed models of society. Usually social studies were limited to small groups of people and were concentrated on researching personalities and behavior. Recently a new aspect of studying social groups, which is more concerned, with dynamical properties of social systems comprised of large numbers of people has been the interest of mathematical sociologists, computer scientists and also physicists. Physicists offer a unique contribution to the study of such systems where the details of individuals are not important but rather how they interact among themselves and how it affects the dynamics of a system and its macroscopic properties. Such approaches have long been present in physics in the field of statistical mechanics.

Network theory provides a simple and very useful framework for studying complex social(and also nonsocial) systems. For the past 200 years, the study of networks was mostly the work mathematicians, with the beginnings stretching back to Leonhard Euler with his solution of the famous Königsberg bridge problem in 1765. Social scientists were the first to conduct empirical studies of networks, first of which was conducted in the 1930s.

One of the first and most famous empirical results concerning social networks was obtained by social psychologist Stanley Milgram at Harvard University, who conducted an experiment in which participants were asked to pass letters hand-to-hand from acquaintance to acquaintance until it reached a designated person. Milgram discovered that of the letters that reached their targets only an average number of six(which was roughly log(N), where N is the total number of people in the USA) people were required to reach the target. This astonishing discovery was found to be present in many social and nonsocial networks and was dubbed the "small world effect".

Although networks have until recently been the interest of mathematicians and social scientists, physicists have contributed fundamental research in the past decade. The reason such research has gained popularity in the past 10 years is probably due to the increased availability of accurate and substantial network data sets.

Unlike mathematicians, who emphasize theoretical models, physicists' researches and theoretical development are founded on empirical results from real-world networks such as World Wide Web studies[4], friendship and scientific collaboration networks etc. Another fundamental aspect that physicists have to offer is the interest in statistical properties of large networks rather than the properties of individual nodes, which are more of interest to social scientists. The motivation for such a perspective comes from statistical physics where a wide array of complex systems has been shown to exhibit universal properties and laws. For instance, magnetism and phase changes in liquids display identical features(e.g. the same scaling factor) even though the systems are qualitatively different. The only thing that governs these universal laws are broader aspect properties of these systems such as the dimension and whether the forces are short or long ranged. Details of the system are not responsible for its overall behavior. Physicists are certain that the case is the same with social systems, the study of which was believed to be more suited from the aspect of complexity theory, where simple rules and large number of interactions develop complex behavior. On the other hand statistical physics was full of cases where a large number of simultaneous interactions actually resulted in simple overall system dynamics.(An example would be the freezing of water where seemingly chaotic complex mechanics of single water molecules start to work together at 0°C and create a freezing transition.)

With these concepts in mind, physicists have discovered interesting properties of social and nonsocial networks some of which have not been addressed before. These discoveries have stimulated new theories, model development and measures to describe these networks. A wide array of researchers including computer scientists, social scientists, biologists, epidemiologists, mathematicians have contributed to the development of network studies, with physicists recently being the leading contributors to foundational work.

Basics of Network Theory

Mathematically a network is represented by a graph, which is a pair of sets $G = \{P, E\}$ where P is a set of nodes $P_1, P_2, ..., P_n$ and E is a set of edges or lines that connect two nodes. Therefore, a network is basically a bunch of points referred to as nodes connected by lines dubbed edges (Fig. 1).



Figure 1. A graph with 5 nodes and 4 edges. The set of nodes is $P=\{1, 2, 3, 4, 5\}$ and the set of edges being $E=\{\{1,2\},\{1,5\},\{2,3\},\{2,5\}\}$.

The foundations of network theory were laid by Leonhard Euler in the eighteenth century. The beginnings of graph study were limited to small numbers of nodes and edges. In the twentieth century, graph study started to apply statistical methods, concentrating on larger systems and studying their parameters, which are then compared to probabilistic models to determine their deviation from randomness. The first two major contributors to probabilistic network theory were Paul Erdős and Alfréd Rényi with their series of papers on random graphs (1959, 1960, and 1961). The introduction of their random graph, which will be more accurately presented in the following, proved to be a useful tool for studying networks.

The properties of networks depend on the way links are distributed among nodes. Some of the most common quantities that used for describing networks are presented in the next section.

Characterization of Networks

In order to study and compare various networks we need a set of concepts and quantitive parameters that describe a network's properties.

Degree Distribution : A node degree k is the number of edges that are connected to a certain node. Different nodes have different degrees and the probability of a randomly selected node having k degrees is given by a degree distribution P(k).

The *small world effect* describes the fact that despite the often very large size of networks the distance between any two nodes is short. A distance between two nodes is defined as the number of edges along the shortest path from one node to another e.g., in Fig. 1 the distance between 1 and 2 is one and the distance between 3 and 5 is two. The most widely known occurrence of this effect is Milgram's "six degrees of separation " i.e. the number of acquaintances between two randomly chosen people in the United States is six.

The *clustering coefficient* is the measure of clustering that appears within a network. A cluster is group of nodes each of which is directly connected to another. So clustering is the tendency of a network to from groups in which every member is connected with each other. In other words, it expresses the probability of C being connected to A if A is connected to B and B is connected to C.

Such a property is known within social sciences as *transitivity*. If we focus on one node *i* that has k_i edges that are connecting it to k_i neighboring nodes then if all the nodes were part of a cluster(commonly referred to as a clique) the total number of edges between them would be $k_i(k_i-1)/2$. A clustering coefficient of a node *i* is the ratio between the actual number of edges and $k_i (k_i-1)/2$:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{1}$$

 E_i is the actual number of edges between the nodes. The clustering coefficient of a network C is the average of all C_i 's. A similar definition is given by:

$$C_{\Delta} = \frac{3 \times (number of fully connected triples)}{number of triples}$$
(2)

The clustering coefficient is concept that was first known in sociology by the name "fraction of transitive" triples (Wasserman and Faust, 1994) the reabson probably being in the usage of definition (2) for the clustering coefficient.

There are also alternative definitions of clustering coefficients used in practice [3].

These are the most common observables in real life network studies. Many other characteristics can also be associated with networks(some examples can be found in [2]). The following two are commonly used in social network studies.

Assortativity is the measure for the correlation between degrees of nodes, which share a common edge. A direct measure of such a correlation would be to calculate the average number of degrees $\langle k_{nn} \rangle$ of neighboring nodes of a node with degree k. If $\frac{d \langle k_{nn} \rangle}{dk} > 0$, the correlation is positive and positive assortativity is present in the network. If $\frac{d \langle k_{nn} \rangle}{dk} < 0$ disassortativity is prevalent in the network and a value of zero would mean that there is no correlation. Positive assortativity means that nodes with higher degree are more likely to be connected to other nodes of higher degree, while negative assortativity suggests that nodes with higher degree are more likely to be connected to nodes with lower degree. There are more rigorously defined ways of calculating assortativity that exceed the interests of this paper[2].

Community structure is a feature commonly associated with social networks. We usually find various social networks to contain communities of common interest, religion, profession... For example communities within the scientific collaboration network are composed of scientists belonging to different fields of research i.e. physicists, mathematicians, biologists etc. There are also divisions within a community i.e. high energy, condensed matter, biological physicists, astronomers etc. The detection of communities within networks has been a focus of many researchers in recent times. Statistical mechanics have proven useful as some of these methods utilizing the Potts and Ising models. Further details can be found in [2].

Network Classification

Though rigorous classification has not been the priority of network researchers, there are standard types of networks that are frequently used by network scientists in their studies.

A very fundamental network, which is often used as a benchmark for other types of networks, is a random network. The concept of a random network was first introduced by Erdős and Rényi. They defined it as N nodes being connected by n edges, which are randomly chosen from N(N-1)/2 of all possible connections. An equivalent definition of a random graph is a binomial model, where every pair of nodes has a probability p of being connected. The average number of connections in such a network is pN(N-1)/2. A random network's main characteristic is that it has Poisson degree distribution:

$$P(k) = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$$
 (3)

<k> is the average degree. Such a distribution means that most nodes have a degree near the average and nodes with very different degrees have a very small probability of being in the network.

The probability that a certain node is connected to a neighbor is equal to the probability of two randomly chosen nodes being connected p. This gives us a clustering coefficient[5] equal to:

$$C = \frac{\langle k \rangle}{N} = p \tag{4}$$

The average distances between nodes in random networks are relatively short. It has been shown [6] that the average distance varies as $\langle S \rangle \sim \log(N)$, where N is the number of nodes. Hence, random networks display the small world property.

The opposite of a random network is a regular network. Regular networks have well known predefined structures in which nodes are usually placed in geometrically regular patterns and connected in simple

ways. Examples of such networks are rings, chains, lattices, trees, fully connected graphs (Fig. 2)...



Figure 2. Examples of regular networks. The images from left to right correspond to the ring, lattice, fully connected and tree types of regular networks accordingly.

It is common for regular networks to have high degrees of order and relatively long average distances between nodes.

Regular networks are usually not associated with social networks but we will see that characteristics of social networks lie somewhere in between the two fundamental types of networks.

Networks that show a power degree distribution are called scale-free networks (Barabási and Albert, 1999):

$$P(k) \sim k^{-\gamma} \tag{5}$$

The term scale-free is used as an analogy from phase transitions, where power laws are frequent, and no characteristic scale can be defined. It has been shown that a lot of actual networks show the scale-free property [5]. A network with a power law degree distribution has a higher probability of having nodes with degrees that are several orders higher from the average degree unlike random networks, where the probability of having a degree very different from the average is very low. Consequently, scale-free networks have more "star-like"(see Fig. 3) appearance with larger groups of nodes with small degree that are connected to the common nodes of very high degree called hubs. Such a structure also accounts for short average distance between randomly chosen pairs of nodes.

Another important aspect of a network is whether it is static or dynamic in time. Social networks are of course dynamic but social studies in the past have frequently been limited to analyses of static network information due to lack of longitudinal data. The studies presented in this paper are longitudinal analyses of a dynamic social network.

Some networks can also have directed edges between nodes. An example is a social network where one



person names the other a friend but the latter does not reciprocate.

Figure 3. The difference between random and scale-free networks(graphs). a) The Erdös-Rényi random-graph model(we construct it by laying down N nodes and connecting each pair with probability p. This network has N = 10 and p = 0.2. Since 45 connecting pairs can be formed, we expect the network to contain approximately 9 links.) b) The scale-free model (it can be constructed by adding new nodes(the red one) to an existing network.) c)The Poisson degree distribution of a random network. d) The power law degree distribution of a scale-free network e)All of the nodes in a random network have approximately the same number of edges. f)"Star-like" structures can be found within a scalefree network with their centers being the nodes with very high degree, which are connected to a large number of small degree nodes.

Social Networks

There have been many studies of various social networks in the last decade. From the collaboration network of scientists and movie actors to phone call networks, citation networks, human sexual contacts networks(for details of all of the latter refer to [5]), epidemiology studies, studies of obesity[9], loneliness[11], happiness[10] spreading within social networks etc. The latter three will be presented in detail in the next chapter. Although studies of social networks were made in very numerous and different study contexts, many properties have been found to be in common for most social networks. Most of them displayed power law degree distribution and clustering coefficients of several magnitudes larger than comparable random networks i.e. random networks generated from the same number of nodes and edges.

The small world effect also seems to be always present in social networks as is confirmed by multiple accounts for the "six degrees of separation" rule(i.e. Milgram's experiment) which indicates an average path of about 6 in some social networks.

It has also been shown that most social networks display positive assortativity in contrast to non-social networks [5]. Degree correlation has a noticeable effect on network topology as can be seen in Fig. 4. Positive degree assortativity is only a special case of a more general phenomenon named assortative mixing in which the probability of two nodes being connected depends on some property of the nodes. In the case of assortativity, that property is the degree of the nodes. An especially common occurrence in social networks, assortative mixing has been associated with the age of a person, the language they speak, race, incomes, education etc. In other words assortative mixing means that like nodes tend to connect with each other i.e. communities are formed within the network. So it is not surprising that assortative mixing is a trait of social networks for we are all aware that communities are present in societies. All of the above-mentioned properties that have been observed in social networks have confirmed what sociologists have repeatedly stated namely that people are grouped in clusters, representing circles of friends and acquaintances, which are all connected with each other and only a few weak links to other parts of societies outside their circles.



Figure 4. Effect of degree correlations on a network. The two networks have the same degree distribution but opposite degree correlation. Network a) has positive degree correlations while b) has negative ones. Red nodes represent high degree nodes and yellow nodes have a low degree. High degree nodes in a) are grouped around the core of the network directly sharing connections, while the ones in network b) are scattered through the whole network with a higher concentration at the edges and are directly connected to mostly low degree nodes.

Examples of Social Network Studies

In this chapter, three examples of social network studies will be presented in detail and discussed. All of the researches were conducted by Nicholas Christakis, a medical sociologist and professor at Harvard, and James Fowler, a political scientist and professor at the University of California. They were joined by John T. Cacioppo, a social psychologist and professor at the University of Chicago, in the study of loneliness.

The social network in all three cases were the participants of the Framingham Heart Study (FHS), which was initiated in 1948 and was intentionally meant as public health study to investigate risk factors for cardiovascular disease.

The methods used for network ascertainment and network analysis were the same for all three studies. They will be presented in short in the following for all three studies. Further details can be found in the original articles [9], [10] and [11].

Source Data and Network Ascertainment

The Framingham Hearts Study enrolled a total of 4 cohorts the first dubbed the "original cohort" in 1948 with n=5209 participants. The second so-called "Offspring Cohort" was enrolled in 1971 and consisted of most(n=5124) of the children with their spouses of the original cohort. The "third generation cohort", which included children of the Offspring Cohort was initiated in 2002 and enrolled a total of 4095 participants. A supplementary minority over-sample cohort named the "OMNI Cohort" was also issued to account for the increased diversity of ethnicity since the original cohort was enrolled. A total of 12067 participants of the FHS were included in all three studies. The information of 7 examinations from 1971 to 2003 was taken into account in each case.

The key cohort of interest in all three cases was the Offspring Cohort with its 5124 participants of which only 10 people left the study over the time span that it was studied and no other losses besides death occurred.

Key subjects are termed "egos" in social network analysis. Any person to whom the ego is liked is called an "alter", so in these cases egos were taken from the Offspring Cohort. In each case alters from all of the cohorts (including the Offspring Cohort) were taken into account so a total of 12067 of individuals were included in the network.

To establish the ties* between the individuals information about the Offspring Cohort was taken from archived, handwritten administrative tracking sheets(which were used to identify people close to the participants for purposes of arranging periodic examinations) and entered in a computer. These tracking sheets provided information about parents, spouses, children, siblings and at least one close friend. In addition, detailed information about home addresses acquired from the tracking sheets enabled the calculation of geographic distance between individuals.

The computerized database was used to identify links between participants and longitudinal information was obtained from one examination to the next.

Friendships were studied more carefully because they can be directional relationships, meaning there are different types of friendship that rely on the direction of the social tie. Three kinds of friendships were determined; an "ego-perceived friendship" in which the ego identifies an alter as a friend but he does not reciprocate, an "alter-perceived friendship" is an opposite case where the alter identifies the ego as a friend and a "mutual friendship" where they both perceive each other as friends. The reason for determining these directionalities was the hypothesis that the influence on the ego was dependent on the type of friendship he shares with an alter. The strongest effect was expected from mutual friendships, followed by ego-perceived friendship, followed by alter-perceived friendships. It was speculated that the person making the identification would esteem the other and wish to emulate him. The results proved the importance of such a differentiation in all of the studies.

Statistical Analysis

Networks were graphed with the Kamada-Kawai algorithm, which repositions the nodes to reduce the number of links crossing each other and visually renders the network according to the analyst's objective.

*a tie is any social relationship an individual has with another e.g. relative, coworker, friend, spouse etc.

In the following, the term "property" will be used for any of the three observables from all of the studies (i.e. obesity, happiness, loneliness).

Measures for determining how much properties were expressed at individuals were obtained either with calculating the body-mass index(for the measure of obesity) or from subcomponents of the Center for Epidemiological Studies depression scale(CES-D) for determining loneliness and happiness. More detailed description of measurement techniques can be found in the original articles([9], [10], and [11]). All three studies observed clustering of "like nodes" which means that people were more likely to be connected with other people who have similar properties. To determine whether clustering was due to chance they compared the probability of individuals with similar properties being connected in the FHS network with a 1000 randomly drawn networks in which the network topology and prevalence of the studied property was fixed to be the same as in the FHS network, but the values of the measure of the property were randomly shuffled for each node. If the probability of egos being connected to alters with similar properties was higher in the observed network then it meant clustering is present in the FHS network, which was the case in all three studies. With this procedure, the researchers were also able to obtain confidence intervals and determine how far the correlation between the properties of alters and egos extended in terms of social distance. Interestingly all three cases found that egos were noticeably affected by alters which were three degrees apart(i.e. the friend of friend of a friend).

Three hypotheses were proposed in all of the studies to acclaim for the clustering. The goal was to determine which mechanism prevails in driving the network towards clustering. The first was *induction*, where one person with a certain property causes the other person to gain that property, the second proposed process was *homophily*, where it is speculated that people prefer to be surrounded with individuals that share the same property and the last being *confounding* or *shared environment*, where people experience contemporaneous exposures that cause them to gain that certain property(e.g. the opening of fast food restaurants causes an increase in local obesity, economic downturns lowers the overall happiness of a neighborhood, matriculating students breaking ties with their old friends and school-mates often feel lonely...).

To distinguish among these processes required repeated measurements of the properties and dynamic, longitudinal information about the forming of ties between people in the network and also information about the direction of the ties. As discussed above it was reasoned that the directionality of a friendship should be a factor to the strength of the influence effect.

The basic statistical analysis was performed using longitudinal-logistic regression models of the ego's property as a function of the ego's sex, age, gender, education and the presence of the property in the previous exam and the presence of the property in the alter at the previous and current exam. By including the measure of the ego's property at the previous exam serial correlations in the errors were eliminated and it helped to control for the ego's genetic predisposition and any intrinsic tendency to have the property(e.g. being obese). Measuring the alter's expression of the property at the previous exam controls for homophily. The possibility of any unknown variables or contemporaneous exposures explaining the clustering was assessed by examining how the directionality and type of social

ties affects the association between the alter's and ego's expressiveness of the property. If such effects were responsible then the direction of the relationship should not be a factor.

The key coefficient in the regression models that measures the effect of induction is on the variable of contemporaneous alter's presence of the property. General estimating equation (GEE) procedures were used for multiple accounts of the same egos across exam waves and across multiple ego-alter pairs. An independent working correlation structure was assumed for the clusters. The GEE regression models provided parameter estimates that are interpretable as effect sizes. The mean effect sizes and 95% confidence intervals(CI) were calculated by simulating the first difference in alter contemporaneous expressiveness of the property using 1000 randomly drawn sets of estimates from the coefficient covariance matrix and holding all other valuables at their means. The sensitivity of the results was also explored by numerous analyses. For details about the statistical methods applied refer to the original articles [9], [10], [11] and the corresponding appendixes.

Results:

The Spread of Happiness[10]

Figure 5 shows the largest subcomponent of the FHS network from 1996 and 2000 based on a restricted set of ties of friends, siblings and spouses. The nodes are colored according to the happiness of the

individual to highlight the clustering, which is significantly higher than in a corresponding random network.



Figure 5. The graphs show the largest subcomponent of friends, siblings and spouses of the FHS network from exams 6(centered in 1996) and 7(centered in 2000). Each node is an individual (squares are male and circles are female). Black lines between nodes mean that the two nodes are siblings and red lines indicate spouses and friends. Colors of the nodes are according to the mean values of the happiness of an ego and the directly connected alters, with yellow color depicting happiest nodes, green means intermediate happiness and blue indicating least happy nodes.

The association of ego and alter happiness was calculated by comparing the probability of an ego being happy if an alter is happy of the measured network with a random network simulated by retaining the existing social ties and prevalence of happiness but with randomly shuffled happiness among the nodes. The results are displayed in Fig. 6. A person is 15.3% (95% CI=12.2% to 18.8%) more likely to be happy if a directly connected alter(a social distance of 1) is happy, 9.8% (7.0% to 12.9%) more likely if an alters' alter is happy(social distance 2) and the effect is 5.6% (2.4% to 9.0%) if alters' alters' alter is happy(social distance 3). The association disappears at four or more degrees of separation from the ego.



Figure 6. The effect of social distance of an alter on an ego's happiness. The probability that an ego is more likely to become happy if an alter becomes happy is displayed. The effect is strongest for alters who are directly connected to an ego and remains significant up to three social distances. The error bars represent 95% confidence intervals.

The association between an ego's future happiness and the total number of happy alters was explored. An ego is 9% more likely to be happy in the future for each additional happy alter connected to him and 7% less likely for each additional unhappy alter.

The simultaneous effect of the total number of alters and the fraction of happy and unhappy alters was also explored. The results showed that happy alters influence the ego's happiness more than unhappy ones and only the total number of happy alters remained the significant factor.

Individual level variables were also studied. As expected the strongest factor for current ego happiness was his happiness at the previous exam. Education, age and sex had the same influence as with previous researches, with women being less happy than men and highly educated people being slightly happier. The examination of the association between different types of alters and their happiness with an ego's happiness was of key interest. The results are presented in Fig. 7.





Using these results, the impact of switching a social contact from being unhappy to being happy was studied. Friends who

live nearby(within a mile~1,6 km) increase the probability of an ego being happy by 25%(CI=1% to 57%), while distant friends(who live more than a mile away) had no significant effect on ego happiness. In addition, the effect of different types of friendships mentioned before on an egos happiness was determined. Nearby mutual friends have the strongest effect 63%(12% to 148%) and alter perceived friends had the least significant effect 12%(-13% to 47%). If confounding was the reason for clustering than the type of friendship should not be relevant in these results.

The effect of other kinds of alters was found to be similar. If coresident spouses became happy the probability of an ego becoming happy was increased by 8%(0.2% to 16%), while noncoresident spouses had no significant effect. Nearby living siblings(less than a mile away) increased the chances by 14%(1% to 28%) and distant(more than a mile away) had no influence consistently.

If next-door neighbors became happy the probability was increased by 34%(7% to 70%), while neighbors living further away in the same block(less than 25 meters away) had no significant effect.

The effects on an ego's happiness were strongly dependent on physical proximity. This suggests that frequent social contacts and not deep social connections were more influential. Interestingly the effects of coworkers were not significant in any case. This suggests that social context can moderate the spread of happiness.

Figure 8 shows the dependence of the probability of an ego becoming happy if an alter becomes happy for friends who live at different physical distances.



Figure 9. The effect of time between alter and ego exams on the ego's probability of becoming happy if alter was happy at the previous exam. The probability decreases with increasing the time.



Figure 8. The influence of friends who live at different distances from an ego on his happiness. Size of the effect declines with increasing distance.

The likelihood of an ego becoming happy increases by 42%(6% to 95%) if friends living less than half a mile away become happy. For friends living within two miles the effect drops to 22%(2% to 45%) and declines quickly at greater distances becoming insignificant for distances greater than 10 miles. The changes of happiness have been reported to be temporary by people getting used to good or bad fortune after some time. This effect can be measured with the association of an alter's previous happiness with the ego's present happiness. These effect sizes are displayed in Fig. 9 as a function of the time between alter and ego exams.

An ego is 45%(4% to 122%) more likely to become happy if a friend examined in the past half year becomes happy and the effect size decreases to 35%(6% to 77%) for friends who were examined in the past year. The effect size rapidly falls for greater periods of time.

Sex was also found to be a significant factor in the spread of happiness. Results showed that happiness spreads easier in same sex relationships than in opposite sex relationships. This could be the reason why neighbors have greater effect on egos than their spouses have.

Socioeconomic status could not be associated with the clustering of happiness for next-door neighbors have a far more significant effect that other neighbors still living in the same block(which have similar wealth and environmental exposures). The geographical distribution of happiness was not related with local levels of income and education. This eliminates the possibility of contextual effects being responsible for the clustering.

The Spread of Loneliness[9]

Figure 10 shows the largest subcomponent of siblings, spouses and friends of the FHS network with nodes colored according to the loneliness of individuals. It shows the clustering of very lonely(blue) and moderately lonely(green) nodes especially at the periphery. To determine whether clustering was by chance the probability of an ego being lonely if an alter is lonely was compared with the probability with randomly generated networks in the same way as in the research from the previous section with randomly shuffling loneliness of the nodes in this case.

The nomenclature used in this study is slightly different from the previous one with egos being referred to as focal participants(FP) and linked participants(LP) being alters. These terms are used in the following charts and graphs.



Figure 10. Largest subcomponent of friends, sibling and spouses(1019 individuals) of the FHS network. Each node represents a participant in the FHS with circles meaning female participants and squares for male participants. Node colors are attributed by the mean number of days an ego and all directly connected alters are lonely. Blue nodes depict very lonely individuals(lonely more than 3 days a week), green ones moderately lonely individuals(lonely two days a week) in yellow node represent persons that are lonely 0-1 days per week. Red lines connect siblings and black ones represent friends or spouses. Clustering of loneliness can be seen and a connection between being peripheral in the network and feeling lonely, which were both, confirmed with statistical models in the research.

The results were consistent with the happiness research(Figure 11). An egos loneliness and an alters loneliness are significantly associated up to three degrees of separation.



Figure 11. The association between an ego's(FP's) and an alter's(LP's) loneliness for different social distances. The relative increase of an ego being lonely given an alter at a certain social distance is lonely was calculated for three exams. Directly connected alters have the strongest relation with the egos and the effect is significant up to three degrees of separation(social distance 3). Error bars represent 95% confidence intervals.

If a directly connected(distance 1) alter is lonely the ego is 52%(95% CI = 40% to 65%) more likely to be lonely. The effect size decreases to 25%(14% to 36%) for alters at two degrees of separation(distance 2) and 15%(6% to 26%) for alters at three degrees of separation(distance 3). The effect virtually vanishes 2%(-5% to 10%) for four degrees of separation.

The analyses of models showed that the most prominent variable was as expected an egos loneliness at the previous exam and showed that ego loneliness is closely connected to the loneliness of their friends. Individuals with more friends felt less lonely with each extra friend reducing the frequency of feeling lonely for 0.04 days per week(which meant two days less of loneliness per year(52 weeks) at an average of 48 days of feeling lonely from the FHS data. That is about a 10% decrease in loneliness for an average person). The same model reported that family members did not have a significant effect.

Loneliness seemed to spread easily than nonlonilness. The probability of an ego becoming lonely if an alter was lonely was two and a half times greater than the probability of a lonely ego becoming unlonely if his alters were not lonely.

Analyses also concluded that loneliness has an influence on network evolution. People that felt lonely compared to those who never do lost on average 8% of their friends in 4 years, the time between exams. Not surprisingly, a person's loneliness did not affect the number of family ties. These effects are symmetric for both incoming and outgoing ties. Loneliness seems to be both the consequence and reason for being disconnected from the network. Lonely people tend to name fewer people as friends and fewer people name them as friends. Our emotions and our social network seem to create a rich-gets-richer cycle, where people that have fewer friends feel lonelier, consequently attract fewer social ties, whereas people with more friends rarely feel lonely, and tend to make more friends over time.

A significant effect on an ego's loneliness was associated with the fraction of lonely individuals connected to that ego(Figure 12). On average individuals surrounded with, lonely alters experience an extra quarter day of loneliness per week compared to those who aren't connected to anyone lonely.



Figure 12. An ego's loneliness at the next exam as a function of the fraction of lonely alters (who reported feeling lonely more than once a week). The fraction of lonely alters is positively associated with the number of days an ego will feel lonely at the next exam. The dotted lines represent 95% confidence intervals.

The influence of different types of social connections on an ego's loneliness was also explored. The results are summarized in Fig. 13.

Each extra day a nearby friend(living within 1 mile of the ego) feels lonely it increases the average number of days the ego feels lonely for 0.29(0.07 to 0.50). Friends who live further away do not have a significant effect(which declines further with greater physical distance). To determine whether confounding effects were relevant factor for loneliness clustering the effect of different types of

friendship were studied. As expected mutual friends were the most influential adding 0.41(0.14 to 0.67) extra days an ego feels lonely per week for every extra day they feel lonely. Next-door neighbors increase the frequency of an ego's loneliness by 0.21(0.04 to 0.38) days per week for every extra day they feel lonely. The neighbors' influence disappears quickly with distance already becoming insignificant for neighbors that live in the same block(within 25 meters). Coresident spouses add an extra 0.10(0.02 to 0.17) days of an ego feeling lonely per week for every extra day of loneliness they experience, while noncoresident spouses do not have a significant affect. Siblings on the other hand do not affect each other's loneliness. These effects suggest that physical proximity is an important factor and consequently the frequency of social interactions plays a key role in the spread of loneliness. Given that siblings have a negligible affect, the social ties that adults chose are more important than those they inherit. The influence of spouses lies somewhere in the middle and contributes less to the spread of loneliness than friends do. These conclusions are consistent with the research presented in the previous section.



Figure 13. The effect of different types of social ties on an ego's loneliness for each extra day the alter feels lonely. Friends, neighbors and spouses have a significant influence but only at close physical distance. Mutual friends have the strongest affect.

The role of a person's gender was also examined. The results showed that women are more likely to be affected by the loneliness of their friends and neighbors and their loneliness is more likely to spread to other people. The possibility of the different baselines in men and women changing the results of the linear model used to derive these results was eliminated by including the sex variable in the regression model, which did show a greater baseline for women.

An individual's depression was also considered as a possible hidden variable that significantly influences the result. The depression index was calculated using all of the questions on the CES-D questionnaire except the ones used for determining the loneliness. A significant correlation between an ego's current loneliness and his current depression was found by adding the variables of contemporaneous and lagged depression of the ego and alter. Adding these variables however did not have a noticeable effect on the association between alter and ego loneliness.

The Spread of Obesity[9]

Figure 14 shows the spread of obesity in a subcomponent of the FHS network from 1975 to 2000. An animation of the evolution of the largest subcomponent of the network is available with the online article(http://content.nejm.org/cgi/content/full/357/4/370/DC2) which shows the spread of the obesity epidemic over the 32-year duration of the study. To explore whether clustering was appearing due to chance the same method of comparing the probability an ego is obese given an alter is obese with a 1000 randomly generated networks with the same topology and obesity prevalence but with randomly shuffled values of the BMI for each node. The probability of an ego being obese if a directly connected alter(social distance 1) was obese was 45% higher in the observed network than in a random network. If the obese alter was 2 social distances away the probability was 20% higher and about 10% higher for three degrees of separation(social distance 3). The effect was negligible for four degrees of separation. The role of an alter's geographic distance from the ego was also explored. While decreasing social distance had a significant effect on the influence of alters' obesity on an egos' obesity, changing geographic distance had no effect. All of the above results are summarized in Fig. 15. To account for homophily a time lagged variable of an alter's obesity was measured and included in the linear regression models. Contemporaneous effects were evaluated by examining how different types of social relationships affect an ego's obesity. To distinguish from confounding effects different types of friendships were also observed(as in the studies in the previous sections). Figure 16 shows the size of the alter's affect on an ego's obesity for different types of social relationships.



Figure 14. The spread of obesity of a part of the FHS network from 1975 to 2000. Each participant is depicted by a circle(node), with red borders representing women and circles with blue borders denoting men. The size of the circle is proportional to a person's body-mass index(BMI). Color are attributed according to an individual's obesity: yellow colors indicate an obese person(BMI \geq 30) and green nodes represent nonobese persons. The colors of ties between nodes depict the type of relationship between individuals: purple lines represent a friendship or spouse and orange ones are family ties. The disappearance of circles means an individual died during these years and the disappearance of a tie means the relationship between two people no longer exists.



Figure 15. Effect of social and geographic distance on the probability of an ego's obesity. Effects were measured at 7 examinations.

Chart A shows the mean effect of an obese alter's social proximity to the probability of an ego being obese. Social distance is presented by degrees of separation(1 denotes one degree of separation from the ego etc.).

Chart B shows the mean effect of the geographic distance of alters with one degree of separation from the egos. Geographic distances are ranked into six mileage groups based on the average distance between homes of directly connected alters and egos: 1 denotes 0 miles(the closest living alters), 2 denotes 0.26 miles, 3 denotes 1.5 miles, 4 denotes 3.4 miles, 5 denotes 9.3 miles and 6 denotes 471 miles(the longest distance between an alter and an ego). Varying geographic distance does not seem to have a significant effect. Error bars represent 95% confidence intervals



Figure 16. The effect of different types of social relationships on the probability of an ego's obesity if an alter becomes obese. The type of friendship has a significant effect on the increase of the risk of an ego's obesity. Mutual friends have the strongest influence. Error bars show 95% confidence intervals.

If an ego and alter were mutual friends than the risk of the ego becoming obese was increased by 171%(59% to 326%) if the alter became obese. The effect size dropped down to 57%(6% to 123%) for ego perceived friends and there was no significant relationship between the ego's and alter's obesity in alter perceived friendships.

Alter's and ego's sex also played an important role. The risk of an ego's obesity increased by 71%(13% to 145%) if a same-sex friend became obese. On the other hand, there was no statistically significant association between the obesity of opposite-sex friends. A man had a 100%(26% to 197%) chance of becoming obese among same-sex friends if the alter became obese, while a woman's risk of becoming obese if her female friend became obese increased by only 38%(-39% to 161%).

In the case of sibling relationships between adults an ego's chances of becoming obese increased by 40%(21% to 60%) if the alter became obese. The effect was also influenced by the alter's and ego's gender. Consistently the effect was stronger between same-sex siblings(55%;26% to 88%) and only 27%(3% to 54%) among siblings of the opposite sex.

The situation was different with familial ties. Between brothers an ego's risk of becoming obese was higher by 44%(6% to 91%) if his brother became obese, while the effect size increased to 67%(27% to 114%) between sisters.

The role of sex was not as important in the case of married couples the ego had a 37%(7% to 73%) higher probability of becoming obese if the spouse became obese and both husbands and wives affected their spouse similarly(44% and 37% respectively).

Interestingly immediate neighbors did not have a significant influence on the egos.

An extra factor of smoking behavior was examined to determine if it had a substantial influence on the results. A measure of the alter's and ego's smoking behavior was added at both the current and previous exams. The coefficients of the alters' influence on the egos were not affected which means smoking did not significantly influence the spread of obesity.

Discussion

Each of the studies showed that people with similar properties tend to cluster within social networks. Although none of the presented researches can provide actual causal mechanisms for the spread of certain properties in social networks a lot of information has still been obtained that enables us to eliminate some previous speculations on these mechanisms and that give an insight on how these mechanisms should look like.

Interestingly there are quite a few characteristics of network behavior that were common in all three studies. In all of the cases different types of friendship between alter and ego were responsible for different sizes of effects an alter had on an ego. This indicates that confounding effects(shared environment, genetic predisposition...) were not important, for such factors would not be affected by the directionality and type of friendships. Also control for the presence of the alter's property at the previous exam eliminated the possibility of homophily being the main contributor to the clustering in the FHS network. Thus in all three cases induction(which can be interpreted as contagion) seems to be the main mechanism behind the forming of clusters of like individuals. The fact that mutual friends were most influential in all three studies also confirms the speculation that esteemed friends have the biggest influence on an individual. The induction hypothesis is further supported by the fact that same-sex relationships were more influential than opposite-sex ones. This agrees with the nature of induction processes, because it seems more likely that people will be influenced more by those they resemble. In addition, social distance was an important factor in each of the researches. A "three degrees of separation rule" was observed in all three studies, which means that an ego's property is significantly associated with the alter, the alter's alter and the alter's alter's alter having or not having that property. Further details behind mechanisms driving the network evolution can be extracted from the results. For

example, the facts that smoking behavior trends and geographic distance do not play an important role in the spread of obesity suggest that norms are relevant to obesity progression rather than behavioral imitation. The latter depends more on the frequency of social contacts, which is strongly connected with the geographic distance between two people. The study of the spread of happiness concluded that happiness was likelier to spread through the network than unhappiness, while the study of loneliness showed that loneliness spreads two and a half times more efficiently than nonloneliness. This difference suggests that although it has been shown that most negative exposures typically have stronger effects than positive ones(Cacioppo & Gardner, 1999), it seems that differential exposure to happy and unhappy events is a bigger factor concerning happiness. So happiness can spread faster through a network, because people are more frequently exposed to friends expressing happiness rather than unhappiness. Another difference observed in the latter two studies is that gender was not an equally important factor. The spread of happiness did not seem to be substantially different among men and women, while loneliness was shown to spread much more easily among women. Such an effect can be accounted for by the fact that women are more likely to engage in intimate disclosures, which would expose their social contacts to expressions of loneliness, while expressing happiness is not predetermined by a trait, which would be more prominent in either men or women.

The fact that social network processes affect an individual and vice versa can be relevant in public health and epidemiological issues. With more accurately targeted network interventions, such as public health interventions, these can become much more cost-effective by targeting key groups of people, which are most effective at spreading trends through the network. An example would be an intervention to reduce loneliness in a network which more aggressively targets people at the periphery of the network, for lonely people are clustered at the edges of the network(because being lonely causes them to lose friends and other ties) and transmit loneliness towards the center of the network. By reducing the loneliness at the edges, the spread of loneliness could not be sustained.

Conclusion

Studying social dynamics offers us results that give an insight on both how human behavior influences the evolution of societies and how social networks influence people on an individual level. Such information can prove very useful in planning public health campaigns, anti-epidemic campaigns, improve addiction interventions by modifying an individual's personal network, optimizing work group organization for higher performance etc. Unfortunately, these studies would require large amounts of data about behavioral and communication patterns and human interaction.

However, in recent times it has become possible to collect and analyze large amounts of data about human behavior and social dynamics. Phone companies have records of calls between people that span many years in the past and search sites, such as Google and Yahoo have records of visited sites and opened documents that individuals clicked. Similarly, every time we send an email it is recorded, the same with bank withdrawals that we make and even every use of our transit card is logged. The design of these everyday systems enables the accumulation of massive amounts of data concerning human behavior. The statistical study of such data would yield interesting results of social dynamics and offer improvements in everyday human social functioning. The study of coworkers' dynamics could result in an optimal collaboration between associates that would increase work efficiency. In addition, mobile phones enable us to trace a large number of people's movements and determine their physical proximity. Such data could be useful for epidemiological studies concerning the spread of a disease(e.g. influenza) which depended on physical proximity.

However, making such information publicly accessible for academic research would violate people's privacy. Although the data is collected anonymously, there have been incidents where the sheer quantity of the data provided statistical methods enough power to de-anonymize the information[8]. Therefore, methods must be developed that protect privacy and at the same time still preserve enough data that would be useful for research. Overcoming this obstacle would highly increase the potential of computational social science, making it a powerful interdisciplinary field of research that could potentially benefit society in numerous ways.

References

[1] Newman M., »The physics of networks«, Physics Today, November 2008, 33 - 38

[2]Sen P., arXiv:physics/0605072v1

[3] Ebel H., Davidsen J., Bornholdt S., arXiv:cond-mat/0301260v1

[4] A.-L. Barabási, »The physics of the web«, Physics World 14, 33-38 (2001)

[5] R. Albert, A.-L. Barabási, »<u>Statistical mechanics of complex networks</u>«, *Reviews of Modern Physics*,74, 47-97 (2002)

[6] R. Albert, H. Jeong, A.-L. Barabási, *Nature* 406, 378 (2000)

[7] R. Cohen, K. Erez, D. ben-Avraham, S. Havlin, *Phys. Rev. Lett.* 85, 4626 (2000)

[8] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, <u>Computation Social</u> <u>Science</u> 323, 721-724 (2009)

[9] Christakis N. A., Fowler J.H., »The spread of obesity in a large social network over 32 years«, *New England Journal of Medicine*, 2007;357:370-9.

[10] Fowler, J. H. & Christakis, N. A. (2008a). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, 337, a2338.

[11]Cacioppo J. T., Fowler J. H., Christakis N. A., »Alone in the Crowd: The Structure and Spread of Loneliness in a Large Social Network«, *Journal of Personality and Social Psychology*, 2009, Vol.97, No.6, 977–991

[12]Strogatz, S. H., »Exploring complex networks«, Nature 410, 268-276 (8 March 2001)

[13]Newman H. E. J., Park J., arXiv:cond-mat/0305612v1

[14] A.-L. Barabási , Taming complexity, Nature Physics, Vol 1, 68-70 (2005)