Abstract:

Homophily is the tendency of individuals to interact with other similar individuals. Homophily has been identified as a key principle in social network formation and evolution. This is because homophily effects the structure of networks and thereby influences how information, diseases, and cultures spread through a network. The number of publicly available networks increases as data becomes more and more digitalized. This increases the number of networks that is being studied. With the increased focus on networks science, it becomes even more important to have a good understanding of homophily. This paper investigates the general Pearsons correlation coefficient measure of homophily, using a dataset created by Ivan Smirnov & Stefan Thurner. The analysis and discussion of this paper is focusing on pointing out the dangers assuming that correlation means causation, by using a combination of Bayesian modelling and investigation of data distributions. The paper concludes that correlation isn't a failproof homophily measurement and average Euclidian distance between ego and alters should be used as a measurement of homophily instead.

Keywords: Social networks, Homophily, Average Euclidian distance, directed networks, temporal networks, network evolution, Ego-alter networks.

Contents

Introduction

"Birds of a feather flock together" is an old English proverb describing people's tendency to socialize with other individuals that share similar traits. This tendency of humans to associate with other humans with similar traits is called **homophily**. A plethora of studies investigating the mechanisms of homophily in peer-networks have surfaced in recent years.[1,2] These papers have been researching the effects- and mechanisms of homophily in social networks. Homophily research plays an important role in predicting network formation and network evolution over time. The more precisely the ecological structure of a network can be predicted, the better the information spread of that network can be calculated and altered.

Networks

Network science is a unifying framework that describes different types of systems under one conceptual lens. Networks differ in types and complexity, but they all consist of the same basic features. The most basic network is two points joined together by a line. The points in a network are referred to as *nodes*. These nodes are also described as agents when describing social networks, as a reference to the social agent acting in the real network. The *lines* connecting the *nodes* are called *edges*. An edge can symbolise a diverse array of interaction types depending on the network, e.g., predatory relation, friendship, or transaction of payments. Many systems can be thought of as networks, exemplary networks could be semantic networks in linguistics, predatory networks in biology, neural networks in artificial intelligence and peer-networks in social sciences.[3]

The use of networks as a framework has allowed for new theories and empirical research[4], as the investigating of a network displays the patterns of interactions between the *nodes* of a system. This allows the researchers to understand new information about a system, that may not be transparent when only investigating nodes in isolation. The interaction patterns and behaviour of a system is decided by the network structure, and can't be inferred by non-network information.[3]

The network investigated in this paper is **directional network**. A directed (or asymmetric) network is one with directed edges, e.g., if one agent likes the picture of another agent, then the other agent isn't obliged to return the like. The studied network is a set of edges between unique nodes that is being measured multiple times, this makes the network used in this paper a **temporal network**. Being a temporal network means that each edge has temporal information and allows for studying the temporal evolution of the network.

Social Networks

A Social network is a network that describes interactions between social agents. The social agent can be a person or a group of people. In network terminology the agent is a *node*, and the interaction between these *nodes* become the *edges*.[5]

Studying social networks is an important necessity for a cognitive scientist, as the investigation of social networks highlights the interactions between individual cognition and social structure. The structure of the network dictates the flow of information, culture, and disease. Understanding the complex nature of human social networks would assist in describing the evolution of a cultural heritage such as language.

A famous example of information flow in social networks is Stanleys Milgram, *The small world problem 1967*[6]. It is this experiment that became known as the six degrees of separation. Milgram was interested in documenting the distance between actors in a network. The participants were tasked with sending a letter to a target person in Massachusetts. The participants had to mail the letter to a person they suspected were more likely to know the target person. They could only mail the letter to a person they knew on first-name basis, and had to write their name on the letter, to document how many times it was mailed. The experiment showed that the average participant reached the target individual by resending a letter six times. Although the number six isn't agreed upon as the universal distance between two average nodes in a network, the *small world effect* is generally accepted as; *the average distance* in vertices between *two nodes is typically very small in relation to the network size.*

Another important tendency of organic networks is that they tend to be **scale-free**, including social networks. This means that degree distribution of the nodes must follow some kind of power law. Resulting in a few nodes being highly connected, whilst most nodes are less connected.[7]

The difficulties of quantifying a social tie

An ongoing challenge of social research is the quantification of peer- and friendships ties. There exists an array of methods that quantify peer-connections, but all the methods come with their own problems. The first obvious method is giving a survey each participant where they must name their peers. This would create an egocentric network for each participant. An egocentric network is a network that is defined from a single focal agent's point of view. The focal node would then be the ego, and the connected nodes would be alters. If all the nodes in a set of ego networks are identifiable, e.g., all the nodes in a social network have a full name, then the set for ego networks would be mergeable by node names. A set of ego networks can therefore be transformed into a social network. But self-reports are expensive and time-consuming, and the method is therefore insufficient for investigating the evolution of social networks[8]. Furthermore, self-reported social data has also been found to highly biased by both recency bias and salience bias[9].

Another method of identifying social ties is therefore needed. The increased digitalisation of our society has substantially increased the documentation of digital communication and digital interactions. The increase of documented digital social interactions also means that a plethora of social networks are available for analysis. This allows for investigation of networks with sizes beforehand thought impossible; such as the internet, scientific paper collaborations[5], and social networks for entire campuses[10]. But the use of digital connections as a proxy for social connections is a considerable prior assumption. Classical sociology describes social acts of symbolism as a form of symbolic capital that acts as a proxy for social capital[11], and thereby being peers. Modern investigations of digital networks has likewise found that likes, comments and digital interactions can be used to successfully identify real-world ties[12].

Homophily

Homophily is the tendency of people to associate with other people with similar traits. The homophily effect has been documented for a large set of human traits. These traits include both visual traits such as gender[13], race[13], and genotype[14]. But, homophily also affects relationship structure based on behavioural traits such as drinking, smoking[15], and sexual orientation[16]. But a cognitively interesting aspect of homophily is how it also affect network formation in regard to discrete values like The Big Five[17], happiness[18] and academic performance[1]. Homophily is considered a fundamental organizational principal of human societies.

The homophily surrounding a social agent will affect their social world significantly. Homophily affects the information the actor receives, the attitudes they form and their experiences with other people[2]. Homophily is documented in a range of networks like spouses, friends[1], co-workers[19] and online forums[16].

Homophily has a range of consequences for the affected individuals, some consequences can be logically reasoned. A logical consequence for homophile networks is that cultural, behavioural, or material information that flows through networks will tend to be localized in clusters of similar agents[2]. This is a consequence of distance of social characteristics translating into distance in the social network. Network distance is the shortest path between two agents, and thereby how many intermediary agents connects the two agents. If two agents have very different characteristics, they will generally have a large network distance between them. If information travels stochastically between agents, then by increasing the shortest path between two agents, then the average time that information needs to travel between the two agents will increase too. This will strand certain information in social spaces, and result in sociological effects like social capital and norms.[2]

Another less logically obvious result of homophily in networks is the reduced diversity in ecologically formed groups. People are more likely to pick their peers as collaborators when forming a group to solve a problem. But this also means that people often create homogeneous groups, as their peers generally are affected by network homophily. A variety of research suggests that a homogeneous group is worse at exploring a problem space, and therefore worse at finding the optimal solution. A cognitively diverse group, with different knowledge sets, is better at finding innovative successful solutions [20]. This is hypothesized to be the result of diverse individuals being more likely to sample for hypotheses at different areas of the hypothesis space. A single individual is likely to have biases and incomplete knowledge about the topic and may therefore not be able to imagine all possible hypotheses. But, since biases, culture and information often get stuck in homogenous clusters of a network this problem is diminished by creating a diverse group. A diverse group quickly cancels the noise of mental hypothesis sampling by increasing the chance of members getting 'stuck' at different peaks, and thereby increasing the chance of sampling the global peak of the hypothesis-probability space.[21]

Formation of homophily

Despite the increase of papers investigating homophily in networks it is still unsure exactly how homophily in networks is formed and how it evolves as time progresses. As previously described, homophily is documented for a wide array of traits, that can be grouped in different subcategories – *observable homophily* and *latent homophily*[22]. Observable traits are responsible for *observed homophily*, and consists of traits such as gender, race, location, and cultural expression. Latent traits explain homophily between individuals that are due to unobservable traits such as personality traits, happiness, and academic abilities. When social networks form, and no biases exists in the set of agents, then the first generation of homophily will be based on observed traits. This indicates, in accordance with research, that people become more homophile over time, as they learn the latent traits of new acquaintances.[23]

Analysis and results

Analysis was performed in R (R Core Team, 2021)[24] using the package rethinking (McElreath, 2021)[25] for fitting Bayesian models. The analysis was performed as a response to Ivan Smirnov and Stafan Thurner's paper, as their homophily measurement didn't show the same results as this papers homophily measurement.

Dataset

The dataset used is this paper is the dataset created by Ivan Smirnov and Stafan Thurner for their paper '*Formation of homophily in academic performance: Students change their friends rather than*

performance'[1]. The dataset is a vector of six temporal adjacency matrices that contains friendship edges between students, and corresponding grades for each student at each temporal trimester. The dataset contains information about 655 students at a Russian public school in Moscow. The Russian school was an average public school, that didn't have any ability-grouping. Most students only had a 10-minute walk, while under 5% of students had more than 20 minutes. The students are anonymised but the Smirnov and Thurner paper informs that the students range from 5th grade to 11th grade. The average cohort size for students in 5th to 9th grade is 108, and the average gender distribution is 44% are girls and 56% are boys. For the 10th and 11th grade the average cohort size is 56 and the gender distribution is reversed; 56% girls and 44% boys.

The academic ability data $G_i(t)$ is a vector consisting of the GPA of student i at time t. The data was collected for the three trimesters of the academic year 2014/2015 and two trimesters of the academic year 2015/2016. It was assumed that the missing spring trimester had the same GPA as before spring vacation, totalling 6 GPA measurements. The academic performance data was published by the school. Along with the GPA of the student, the full name was also published.

Creating a peer-network

The full name of the students was used to constructing construct a peer-network. To construct the peer-network the Russian social media VK (http://vk.com) was used. VK is the largest social media website in eastern Europe, with almost 100 million users. Only 5% of students didn't have an active profile. VK provides features that are alike those of Facebook and allows users to form mutual online friendships and interact with each other's content. However, the friendship feature of VK was not used as a peer indicator, as it was to static in the temporal scope, as accepted friendships rarely was disconnected. Instead, friendship was measured by 'likes' on other student's pages. If a student had liked a post from another student during a trimester, they get an outgoing friendship edge to that student. This gives a more ecological and flexible friendship definition that evolves over time. The result is a $N \times N \times \mathrm{t}$ adjacency matrix $A_{\mathrm{ij}}(t)$, where $A_{ij}(t) = 1$ indicates that student i liked a post from student j during the time interval $t - 1$ to t. Where N is the total amount of students, and t is the trimester.

Table 1. Dataset description

Metric	*Total value*
Nodes	3,930
Edges	26,050
Avg. degrees	6.66 (SD = 9.55)
Avg. GPA	3.85 (SD = 0.55)
Avg. homophily	0.64 (SD = 0.28)

Calculating ego-alter homophily for a continuous variable

For each student *i*, $G_i(t)$, a set of ego-alter networks was constructed from the adjacency matrix $A(t)$. The average degrees of the ego-alter networks were 6.6, meaning that each ego, student *i*, on average had almost 7 peers. A homophily measure, H, was then calculated for each network, using an average Euclidian distance between the GPA of the ego and the GPAs of the alters. The average Euclidean distance between ego and alters is defined as

$$H = \sqrt{\frac{\sum_{j=1}^{N}(a_j - e)^2}{N}}$$

Where N is the total amount of alters, j indexes the alters, a_j is the GPA of alter j and e is the GPA of the ego node. Using the Euclidian distance as a homophily measurement gives a homophily measurement for each trimester, for each student. This differs from Ivan Smirnov & Stefan Thurner's paper[1] that uses persons correlation coefficient between the vector of the student's GPA and the vector of the average GPA of student's friends. Pearsons's correlation coefficient only gives a singular homophily measurement for each trimester, that describes the *correlation* between the GPA of the students and their respective friends.

Investigating the differences of homophily indexes

The paper will investigate the two homophily indexes described above. Figure 1(A) shows a violin plot of the mean Euclidian distance between ego-alter GPA. Figure 1b shows the Pearson correlation coefficient between the GPA of the ego node and the average GPA of the alter nodes. Since the Euclidian homophily shows no signs of an increase in GPA-homophily and the correlation coefficient shows signs of homophily, a further investigation of causality and biases was needed.

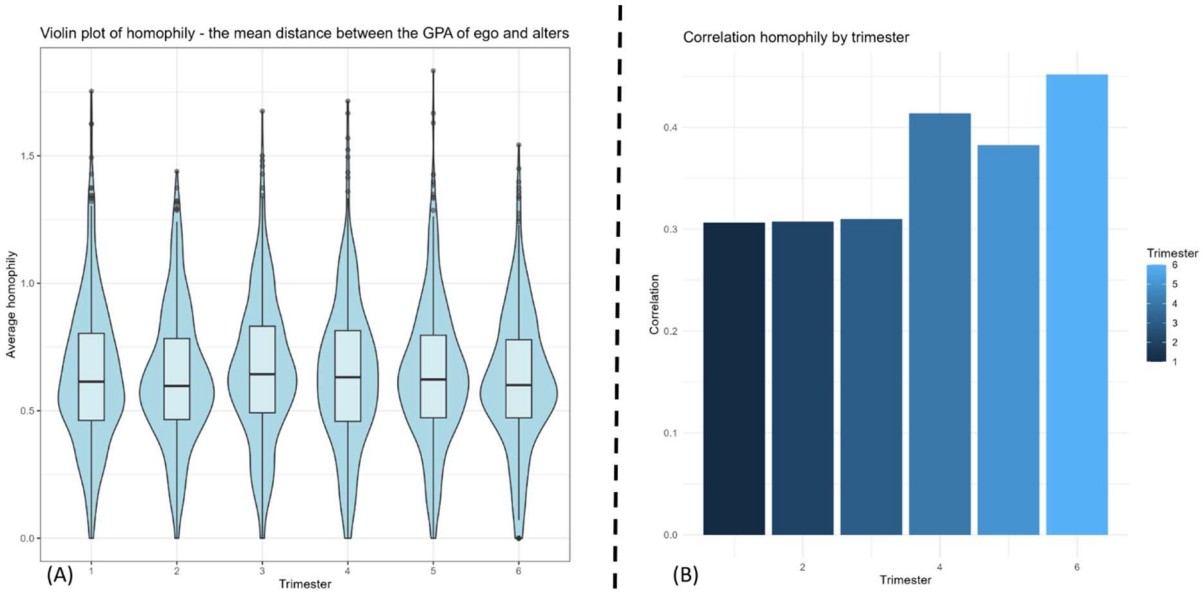


Figure 1(A). Is a violin plot with overlayed boxplot of the Euclidian distance homophily measure for each trimester. Figure 1(B) depicts the average persons correlation coefficient homophily measurement for each trimester.

To ensure that the average Euclidian distance doesn't have any unseen trends, it was investigated with two models. The first model is used to determine if there is an increase in mean Euclidian homophily. The models' prior- and posterior distributions are as following:

Simple model for homophily:

$$Homophily \sim Normal(\mu, \sigma)$$
$$\mu = \alpha + \beta_T * Trimester$$
$$\alpha \sim \text{Normal}(0, 0.5)$$
$$\beta_T \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

Posterior distribution of parameters

	Mean	SD	n_eff	rhat
α	0.66	0.02	1282	1
β_T	0.00	0.00	1306	1
σ	0.26	0.00	1754	1

Figure 2. Priors for the simple model of Euclidian homophily and posterior distributions calculated using MCMC sampling.

A secondary multi-level model was constructed to ensure that the increase in correlation coefficient homophily wasn't caused by patterns in individual-level effects. This was done by including student ID as a parameter:

Multilevel model for homophily:

$$Homophily \sim Normal(\mu, \sigma)$$
$$\mu = \alpha + \beta_T * Trimester + \beta_{ID} * ID$$
$$\alpha \sim \text{Normal}(0, 0.5)$$
$$\beta_T \sim \text{Normal}(0, 0.5)$$
$$\beta_{ID} \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

Posterior distribution of parameters

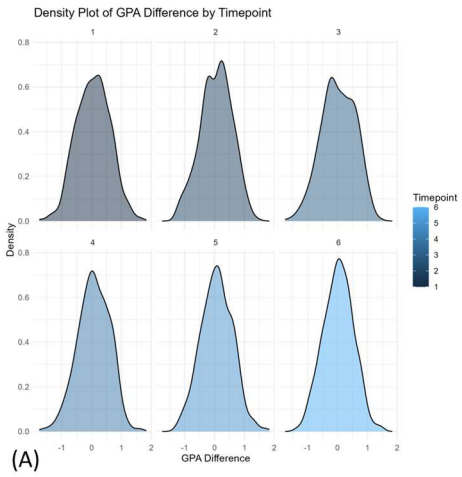
	Mean	SD	n_eff	rhat
α	0.33	0.27	8	1.80
β_T	0.00	0.00	1306	1
β_{ID}	0.00	0.00	8	1.38
	...	...	...	...
	0.37	0.28	12	1.64
σ	0.19	0.00	395	1.01

Figure 3. Priors for the MLM of Euclidian homophily and posterior distributions calculated using MCMC sampling.

As none of the homophily models indicates that Euclidian homophily for academic performance increases over time, a further investigation of the correlation index is necessary. As the correlation homophily index is defined as $\text{corr}(G_i(t), G_i(t) + D_{j...N})$, where G_i is the GPA of student i at trimester t, and $D_{j...N}$ is the average *difference* in GPA for direct friends j to N.

This definition indicates that if the average difference in GPA for alter nodes is a gaussian distribution, then the correlation will be defined by the gaussian distribution. As the correlation between a **vector** and a **vector + a perfect gaussian distribution** will be determined by the standard deviation of that distribution, and not the causality of the distribution.

So, the average GPA-difference distribution for each trimester is plotted:



(A)

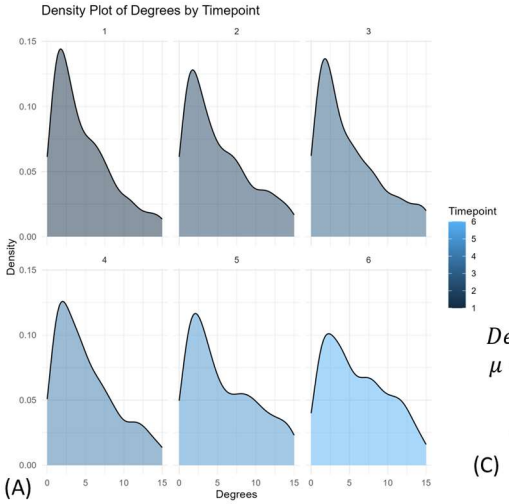
Trimester	Mean difference	SD
1	0.05	0.56
2	0.06	0.53
3	0.03	0.56
4	0.07	0.53
5	0.08	0.53
6	0.03	0.5

(B)

Figure 4(A). Density plots for GPA difference between student and direct friends for each trimester. Figure 4(B). A table describing the distributions for each trimester, assuming they are gaussian.

A possible trend in the GPA-difference distributions seen in figure 3(A) can be visually investigated. It seems that as the trimesters (t) increases the GPA distribution becomes unimodal, rather than bimodal, and becomes more symmetrically distributed around the mean. If the GPA distributions approximates a gaussian distribution more closely as the trimesters increase, then the increase in causation could be explained as result of increased normality. A decrease in standard deviations would also explain the increased correlation.

So, to understand the distributions increasing approximation of a gaussian distribution, the causation of the increased normality of the distribution needs to be investigated. The number of ego network with alters created for each trimester is relatively static ($\mu = 462, \text{sd} = 25$), but the sizes of ego networks are not, as seen in figure 4(A). The average amount of degrees for each ego-network increases with the trimesters. If the students on average samples more students that are alike them as peers, the chance of sampling the the correct mean of their ideal peer population increases.



Trimester	Mean degrees	SD
1	4.71	7.08
2	6.10	8.98
3	6.56	9.17
4	6.24	9.82
5	8.23	11.08
6	7.93	10.25

(B)

Model for degrees:

$$Degrees \sim Normal(\mu, \sigma)$$
$$\mu = \alpha + \beta_T * Trimester$$
$$\alpha \sim \text{Normal}(0, 0.5)$$
$$\beta_T \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

(C)

Distribution for parameters

	Mean	SD
α	4.03	0.33
β_T	1.33	0.09
σ	10.16	10.16

Figure 5. (A) Density plot of average degrees of ego-networks as trimesters increase. (B) Table describing the degree distributions for each trimester. (C) Model describing avg. degrees as a distribution that depends on trimester as a parameter with corresponding posteriors from MCMC sampling.

A Bayesian model was also created to describe the relationship between trimester and the average amount of alter nodes (degrees), the model and posterior distributions is shown in figure 4(C). The prior distribution of degrees was a normal distribution to minimize assumptions.

A last sanity check was done, by creating a dataset where all GPAs was randomized before sampling average GPA of friendships. For this dataset, the average correlation homophily was 0.

Discussion

This paper was meant to investigate a conditional model of homophily but was halted by finding different results than Smirnov & Thurner's paper. Instead, the focus of analysis in this paper was then changed to showing that '*correlation isn't causation*'. This was necessary as the dataset didn't contain the homophily that could be used to model homophily of academic performance, but rather contained spurious correlations that looked like homophily of academic performance.

As seen in the violin plot of figure 1(A) the average distance of an arbitrary students GPA and the GPA of that students' direct friends didn't change as time increased. Even though some degree of homophily was present at the first trimester, there was no increase in Euclidian homophily in contradiction to what the correlation measurement in figure 1(B) shows. If the correlation increases without the distance in GPA decreases, then that correlation is caused by another factor than finding friends with a more alike GPA. Smirnov & Thurner argued that because the correlation increased over time when they fixed the GPA of students at the value of their GPA in the first trimester, then students must exchange their peers with new peers who are more alike in the GPA-dimension. But as the average distance in homophily stays the same, it isn't the GPA-dimension the students use to align themselves.

However, the increase in correlation isn't purely a result of an increase of degrees either. Because if that was the case, then we would see the same trend in correlation increasing for the dataset where students had randomized GPAs. Just increasing the number of degrees of the average student's ego network would result in a biased sample because of it being a scale-free network. The student with the most incoming directed edges would then be overly sampled and ruin the normality.

This means that a third theory would have to be the causation of the hidden correlation. A possible theory could be: As the degrees of the average student increases, that student has more possibilities of creating friendships with students with similar observable traits, causing observable homophily. If these traits all differently influence the GPA of a student. Then sampling friends with relatively similar traits would allow the student to sample a distribution of friends with relatively similar GPAs to themself. The mean difference of the friends' GPAs would then centre around zero as the sample size of alike students increases. But the average Euclidian distance of GPAs in these samples would stay the same, as that would be caused by the sampling error of variations in observable traits. However, investigating this would require a social dataset with multiple observable- and latent traits, and is therefore outside the scope of this paper. But, using the average Euclidian distance in an ego-alter network as a homophily index would scale well into a multidimensional social space like described above.

Homophily is an important factor of social networks, but also a factor with unseen spurious associations. Future investigations into the nature of homophily is necessary but should be done carefully, and by a

diverse group of scientists. But work is still needed before homophily research is ready to leave its nest and help the field of network science take flight. Future research should address the multi-dimensionality of social life further and study how it affects social network structures. But until then, *birds from a feather will continue to flock together*.

References

1. Smirnov I, Thurner S. Formation of homophily in academic performance: Students change their friends rather than performance. Masuda N, ed. *PLoS ONE*. 2017;12(8):e0183473. doi:10.1371/journal.pone.0183473

2. McPherson M, Smith-Lovin L, Cook JM. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*. 2001;27(1):415-444. doi:10.1146/annurev.soc.27.1.415

3. Newman M. *Networks*. Oxford University Press; 2018.

4. Baronchelli A, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N, Christiansen MH. Networks in Cognitive Science. *Trends in Cognitive Sciences*. 2013;17(7):348-360. doi:10.1016/j.tics.2013.04.010

5. The structure of scientific collaboration networks | PNAS. Accessed August 1, 2023. https://www.pnas.org/doi/full/10.1073/pnas.98.2.404

6. Travers J, Milgram S. An Experimental Study of the Small World Problem**The study was carried out while both authors were at Harvard University, and was financed by grants from the Milton Fund and from the Harvard Laboratory of Social Relations. Mr. Joseph Gerver provided invaluable assistance in summarizing and criticizing the mathematical work discussed in this paper. In: Leinhardt S, ed. *Social Networks*. Academic Press; 1977:179-197. doi:10.1016/B978-0-12-442450-0.50018-3

7. Sciam.2003.scalefree.pdf. Accessed August 8, 2023. http://compbio.korea.ac.kr/wiki/images/d/d4/Sciam.2003.scalefree.pdf

8. Wuchty S. What is a social tie? *Proceedings of the National Academy of Sciences*. 2009;106(36):15099-15100. doi:10.1073/pnas.0907905106

9. Inferring friendship network structure by using mobile phone data | PNAS. Accessed August 7, 2023. https://www.pnas.org/doi/full/10.1073/pnas.0900282106

10. Empirical Analysis of an Evolving Social Network | Science. Accessed August 7, 2023. https://www.science.org/doi/full/10.1126/science.1116869

11. Bourdieu P. THE FORMS OF CAPITAL.

12. Inferring Tie Strength from Online Directed Behavior | PLOS ONE. Accessed August 7, 2023. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0052168

13. Shrum W, Cheek NH, Hunter SMacD. Friendship in School: Gender and Racial Homophily. *Sociology of Education*. 1988;61(4):227-239. doi:10.2307/2112441

14. Correlated genotypes in friendship networks | PNAS. Accessed August 2, 2023. https://www.pnas.org/doi/full/10.1073/pnas.1011687108

15. The contribution of influence and selection to adolescent peer group homogeneity: The case of adolescent cigarette smoking. Accessed August 2, 2023. https://psycnet.apa.org/record/1995-04997-001

16. Thelwall M. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*. 2009;60(2):219-231. doi:10.1002/asi.20978

17. Selden M, Goodie AS. Review of the effects of Five Factor Model personality traits on network structures and perceptions of structure. *Social Networks*. 2018;52:81-99. doi:10.1016/j.socnet.2017.05.007

18. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study | The BMJ. Accessed August 2, 2023. https://www.bmj.com/content/337/bmj.a2338

19. Ibarra H. Homophily and Differential Returns: Sex Differences in Network Structure and Access in an Advertising Firm. *Administrative Science Quarterly*. 1992;37(3):422-447. doi:10.2307/2393451

20. Lungeanu A, Contractor N. The Effects of Diversity and Network Ties on Innovations: The Emergence of a New Scientific Field. *American Behavioral Scientist*. 2014;59:1-17. doi:10.1177/0002764214556804

21. Bang D, Frith CD. Making better decisions in groups. *Royal Society Open Science*. 2017;4(8):170193. doi:10.1098/rsos.170193

22. Latent Homophily or Social Influence? An Empirical Analysis of Purchase Within a Social Network | Management Science. Accessed August 7, 2023. https://pubsonline.informs.org/doi/10.1287/mnsc.2014.1928

23. Block P, Grund T. Multidimensional Homophily in Friendship Networks. *Netw Sci (Camb Univ Press)*. 2014;2(2):189-212. doi:10.1017/nws.2014.17

24. R Core Team (2020). — European Environment Agency. Accessed January 4, 2023. https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006

25. Statistical Rethinking | A Bayesian Course with Examples in R and Stan. Accessed August 7, 2023. https://www.taylorfrancis.com/books/mono/10.1201/9781315372495/statistical-rethinking-richard-mcelreath