Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper PUBLISHERS key to the Researchers

Editorial
Volume 1 Issue 1 - January 2017
DOI: 10.19080/BBOAJ.2016.01.555551

Biostat Biometrics Open Acc J

Social Network: A New Paradigm for Modeling Human Interaction: Implications for Statistical Inferences

Chen T[1]*, Lu N[2], White AM[3], He H[4], Wu P[5], Hui J[6], Feng C[6], Zhang H[7], Kowalski J[8] and Tu XM[6]

[1]*Department of Mathematics and Statistics, University of Toledo, USA*

[2]*Department of School of Medicine and Health Care Management, Huazhong University of Science and Technology, China*

[3]*Department of Psychiatry, University of Rochester, USA*

[4]*Department of Epidemiology, School of Public Health & Tropical Medicine Tulane University, USA*

[5]*Department of Value Institute, Christiana Care Health System, USA*

[6]*Department of Biostatistics and Computational Biology, University of Rochester, USA*

[7]*Department of Biostatistics, St. Jude Children's Research Hospital, USA*

[8]*Department of Biostatistics and Bioinformatics, Emory University, USA*

Submission: November 26, 2016; **Published:** January 09, 2017

***Corresponding author:** Chen T, Department of Mathematics and Statistics, University of Toledo, Toledo, OH 43606, USA,
Email: tian.chen@utoledo.edu

Abstract

A broad of spectrum of disciplines have adopted social network data to examine relevant contextual issues in a wide array of fields. Yet, statistical methods to address biases in statistical inference introduced by the between-subjects relationship within the context of node, or subject, interaction in social networks are underdeveloped. Traditional statistical models define relationships among measures of within-subject attributes, i.e., measures of attributes from each subject. The between-subject attribute for node (subject) interaction in social networks is both conceptually and analytically different from the within-subject attribute. As a result, conventional statistical methods such as t-test and linear regression models are fundamentally flawed when applied to model between-subject attributes in social network settings. We illustrated fundamental differences of the between- and within-subject attributes and resulting implications for social network data analysis of social network densities. We also proposed a new paradigm to model between-subject attributes and illustrate the approach with the analysis of social network density.

Keywords: Between-subject attributes; Network density; Within-subject attributes; U-statistics

Introduction

Social network analysis (SNA) measures and maps “connectedness” or “flows” (e.g., information, resources, etc.) both within and across individuals, groups and organizations. Network analyses can illustrate how the links or ties among elements of a system affect the outcomes or emergent properties generated when interactions occur between individual and organizational entities. Social network data and analysis provides a new paradigm for social and behavioral sciences by creating a new data dimension of human interaction. With data science methods, recent studies have indicated that human interaction is a key predictor of most human behaviors and social phenomena such as flu pandemics, financial crashes and political upsets Pentland et al. [1]. Indeed, human interaction is such a strong class of predictors that it fundamentally changes the way we design behavioral and social intervention research studies.

For example, in a study using mobile and online social media, Pentland et al. [1] showed that simply changing the schedules of coffee breaks from one person at a time to multiple employees simultaneously resulted in a productivity increase of $15 million a year for a Bank of America call center. Another study about helping save energy found that it is more effective to change behaviors of others connected to the person of interest than to try to change this person in the group who is consuming more energy Pentland et al. [1]. The researchers provided small cash incentives to individuals who had the most interaction with specific high energy use consumers, rewarding them for improved behavior of offending consumers. Similar studies replicate this finding that a social influence approach is up to four times as efficient as traditional methods Pentland et al. [1].

Although computer and computational scientists have led the research and application of social networks, available statistical methods are quite limited and in particular typically only provide statistics that describe observed patterns of interest. Although such descriptive statistics may be sufficient for applications in Bigdata such as online social media where the sample size is so large that sampling variability becomes negligible, they cannot be used for inferences about observed patterns at the population level for most research studies with limited sample sizes. The latter is the hallmark of modern research studies in the biomedical, behavioral and social sciences.

For example, if the sample mean of an outcome such as age in a random sample from a population of interest is 50, it is well known that this mean is an estimate of the average age of all individuals comprising the population, i.e., the population or mathematical mean. A conference interval is generally used to provide an indication for the accuracy of this estimate (errors due to sampling variability) when used to infer the population mean. Available methods for social network data largely focus on calculating quantities that summarize features of the sampled social network. Although some methods and software packages have been proposed for inference, none provides correct results. These packages fail to address the new conceptual as well as analytic challenges induced by the dimension of human interaction.

Take, for example, the social network density, a popular index of connections between subjects in a social network Wasserman & Faust [2]. For a sampled social network, this index is readily computed. However, if we want to use this sample density as an estimate of the mean density underlying the population of the social network of interest, no available method provides correct standard error or confidence interval estimates. In this paper, we discuss major flaws in the current data and analytic paradigm when applied to model social network data. We introduce new data concepts that capture human interaction in social networks and their associated analytic issues. We focus on the social network density and discuss flaws in modeling such a popular index using conventional statistical methods as appeared in the literature. We propose a new approach to address the flaws and illustrate the proposed method using simulation data.

Between-subject vs. within-subject attributes

In the current data and analytic paradigm, outcomes, or variables, are defined as measures of attributes from each individual, such as demographic information, medical illnesses, and social and health behaviors. If we denote each subject in the population of interest as S_i, these outcomes are well defined for each S_i. For example, if $y(S_i)$ denotes suicide attempt (a binary indicator) for the ith subject, this variable, often simply denoted as y_i, is well defined for this subject. Research studies under the current paradigm, regardless of observational or randomized control studies, all focus on modeling relationships among such variables of within-subject attributes. For example, in a study on suicide attempt, one may fit a statistical model such as logistic regression to see if depression, physical illness, major life events, etc., contribute to the occurrence of such a self-harm event. The variables and statistical models here all focus on relationships among variables from each individual, completely ignoring influences from interactions with others.

Outcomes in social networks not only include such within-subject attributes, but also interactions among different individuals. The latter is unique for social network data. Unlike individual outcomes in conventional analysis, interaction between individuals is a between-subject attribute and is not definable based on one individual. To illustrate, consider the social network density, a popular index for network connectivity defined as the mean connection between two subjects in the network. Specifically, if we denote two subjects in the social network of interest as S_i and S_j, the connection between the two individuals is a binary indicator involving both subjects, $I(S_i,S_j)$, which is 1 (0) if S_i and S_j are connected (otherwise).The social network density is the mean of this indicator, $\theta=E[I(S_i,S_j)]$, where $E(.)$ denotes the mathematical expectation. Unlike the outcome in conventional analysis, the outcome here, $I(S_i,S_j)$, must be defined based upon two subjects. As we elaborate in detail in the next section, the change from the within- to between-subject attribute in the social network outcome presents a serious challenge for modeling social network data under the current paradigm.

Note that in social network analysis, sometimes connection may be defined as representing an asymmetric relationship. For example, if we distinguish the nature of initiating and accepting an invitation for a relationship, then $I(S_i,S_j)$ may not be the same as $I(S_j,S_i)$ Cohen & Havlin [3]. For ease of exposition, we only consider symmetric relationships such that $I(S_i,S_j)=I(S_j,S_i)$ unless stated otherwise.

Fundamental flaws in current paradigm for modeling social network data

As noted in the preceding section, conventional statistical analyses focus on relationships among within-subject attributes. The qualifier “within-subject” is critical, since it is the foundation of the current statistical modeling enterprise. Statistical independence is the basis for inference for statistical models Tang, et al. [4]. This fundamental assumption is reasonable for measures of within-subject attributes. In practice, if we draw a random sample from a population of interest, it is reasonable to assume that all subjects sampled do not interact with each other so that the outcomes of interest are stochastically independent. In some studies, certain observations may share something in common and the independence assumption may be invalid. To analyze such data, we must take the potential dependence into consideration and develop methods to address it.

How to cite this article: Chen T, Lu N, White A, He H, Wu P. et.al. Social Network: A New Paradigm for Modeling Human Interaction: Implications for Statistical Inferences. Biostat Biometrics Open Acc J. 2017;1(1): 555551. DOI: 10.19080/ADOH.2016.01.555551

For example, longitudinal data create non-independent observations Tang et al. [4]. In such studies, each subject is repeatedly assessed over time. Since outcomes from the same subject are generally less variable when compared with those from other individuals, independence is not a reasonable assumption when applied to all the outcomes. However, if we view the repeated outcomes from the same subject as a data cluster, then these clusters are stochastically independent. This is precisely the basis for all modern statistical models for longitudinal data Tang, et al. [4]. This partitioning strategy is also used in dealing with dependent observations from other settings, such as groups of individuals living in different geographic and economic regions with respect to the values of the outcomes of interest Tang, et al. [4].

Unfortunately, the strategy of data partitioning deployed in clustered data analysis such as longitudinal data does not work for between-subject attributes in social network analysis. To illustrate, consider again the outcome for the social network density. For notational brevity, denote the connection indicator between two subjects simply by $I_{ij}=I(S_i,S_j)$. For a sample of social network consisting of n subjects, there are a total of $1/2n(n$-1) such variables. These variables are clearly not independent. For example, since I_{12} and I_{13} both involve the first subject, they are not independent. Further, there is no way to partition the $1/2n(n$-1) indicators into independent clusters as in modeling clustered data, because of interlocking relationships among the variables. For example, consider any pair of indicators, I_{ij} and I_{kl}. To see their dependence, simply note their relationships with a third indicator I_{kl} .Since I_{ij} and I_{jk} both involve subject j , they are dependent. Likewise, I_{jk} and I_{kl} are dependent as well. These relationships imply that I_{ij} and I_{kl} are not dependent.

The above analysis shows that the current statistical modeling paradigm is fundamentally flawed when used to model social network data, since stochastic independence is not a valid assumption for the between-subject attribute within the social network setting. Without addressing this fundamental barrier, research capitalizing on key concepts in social networks such as social support or information will be seriously hampered. Unfortunately, conventional statistical models such as regression have been and are continuingly being applied to model social network data [5-12]. Such methods have unpredictable effects and study findings may potentially misinform research communities and the public. In the next section, we propose a new paradigm for modeling variables of between-subject attributes. We motivate and develop the proposed approach by focusing on the network density.

A New Paradigm for Modeling Social Network Outcomes

Independence assumption in traditional paradigm

Social network density is a popular measure for connectedness of subjects in a population of interest. In the nomenclature of social network graph analysis, members of a social network are referred to as nodes. To highlight the distinction between the between- and within-subject attribute within the context of statistical analysis, we use subjects as members of a social network unless stated otherwise. The network density metric is commonly employed in hypothesis formulation and testing involving social network data structures. For instance, in child development, scientists have sought to infer whether the extent to which students in a school system know one another is related to observed health risk behavior prevalence such as smoking or sexual risk-taking Lakon et al. & Ramirez-Ortiz et al. [13,14]. In prevention and health sciences, population-level health (e.g., prevalence of drug use or infectious disease) is examined relative to the degree to which professionals (e.g., social workers and other health professionals) in a network know one another or collaborate [15-17]

Consider a sample of social network consisting of n subjects from a population of interest. The connection indicator I_{ij} is a variable defined for each pair of subjects, with the value 1 (0) if the subjects are connected (otherwise).The social network density is defined as the average value of the indicator at the population level, i.e., $\theta = \mathrm{E}(I_{ij})$.Given the sampled social network, we can estimate by the sample mean,

$$\theta^\wedge = 2/n(n\text{-}1)\Sigma_{1}\leq_{i}<_{j}\leq_{n} I_{ij}, \qquad (1)$$

Where $\Sigma_{1}\leq_{i}<_{j}\leq_{n} I_{ij}$denotes the summation of I_{ij} over $1\leq i<j\leq n$.For example, if n=4 ,

$$\theta^\wedge = 4*3/2\ (I_{12}+I_{13}+I_{14}+I_{23}+I_{24}+I_{34}) \qquad (2)$$

The estimate $\theta^\wedge$ has a different expression than the sample mean of a variable of within-subject attribute, because of potential connections between subjects in the social network. For example, if y_i is a variable of within-subject attribute such as income for each individual in the sample, then the (sample) mean income, $\mu=1/n\ \Sigma 1\leq_{i}\leq_{n} y_i$, where $\Sigma 1\leq_{i}\leq_{n} y_i$ denotes the summation of y_i over $1\leq i\leq n$. In this case, the summation is over the individual variables, rather than the n individual variables y_i, rather than the $n(n$-1)/2 between-subject attributes.

The assumption of independent observations is particularly important when making inference about parameters of interest (e.g., hypothesis testing and confidence intervals). For example, in the case of within-subject attributes such as the variable y_i above, this assumption implies that the y_i 's are independent of each other.As noted earlier, this is clearly not the case with the between-subject attributes I_{ij}. Further, within the context of social networks, even independence regarding the y_i 's becomes questionable as well. For example, if connection measures business relationships, it is quite plausible that connected individuals have more similar incomes than unconnected ones.

Thus, when modeling social network data, even variables of within-subject attributes may become dependent.

Inference for Social Network Density

As discussed in Section 3.1, the density of a population social network is the mathematical expectation $\theta=E(I_{ij})$ of the connection indicator I_{ij}. We estimate θ, the proportion of connected subjects in the population network, by the sample mean $\theta^{\wedge}$ in (1), which is the proportion of the indicators with the value 1 among the $n(n\text{-}1)/2$ indicators in the sampled social network. Although methods for inference about proportions exist, these conventional statistical methods cannot be applied to compute standard errors and construct test statistics, because of lack of independence among the $n(n\text{-}1)/2$ variables I_{ij}. Further, as noted in Section 3.1, even modern statistical models for clustered data do not address this dependent issue, because of lack of independent clusters in the data.

The dependence of I_{ij} is the result of defining the connection indicator on a pair of subjects. This particular type of dependent variables was first investigated by Hoeffding [18]. His pioneering work has led to the development of a systematic treatment of such dependent variables, known as the theory of U-statistics. Classic examples of U-statistics include the Mann-Whitney-Wilcoxon (MWW) rank-sum test, the Wilcoxon signed-rank tests [19] and methods for analysis of receiver operating characteristic curves [19-21]. The key difference between the U and conventional statistic is the between- vs. within-subject attributes.

For example, in the MWW rank-sum test for comparing some outcome of interest between two groups, the between-subject attribute is a binary indicator with the values defined based on whether the outcome of an individual from one group is larger than that of a subject from the other group. Within the current context of social network density, the between subject attribute is the binary indicator of connection status. Note that although the indicators in both examples are a measure of between-subject attributes, they are still different. The one for the MWW test is constructed from comparing individual outcomes (or within-subject attributes), while the one for the social network density is not. In this sense, the connection indicator for the social network density is an intrinsic between-subject attribute, unique to the social network data.

Using the theory of U-statistics, we can evaluate the standard error of the sample network density, $\theta^{\wedge} = 2/n(n\text{-}1)\Sigma_{1}\leq_{i}<_{j}\leq_{n} I_{ij}$, despite the interlocking relationship across the observations I_{i} [21,22].In addition,

following the theory of U-statistics, the sample network density also has an asymptotic normal distribution, allowing us to make inference about the population network density.

A Simulation Study

We conducted a simulation study to illustrate the U-statistics approach and examine the impact of dependence of the between-subject attribute I_{ij} on inference by comparing the proposed approach with conventional and available alternatives.

We considered a sample of social network consisting of n=100 subjects. To create the connection indicator, we first generated two variables Z_{i} and Z_{j} from a standard normal with mean 0 and variance 1 and then defined the connection indicator as: $I_{ij} = I_{\{Zi+Zj\leq 0\}}$. In this special case, the true value of the density is θ=0.5. According to the theory of U-statistics, the sample density $\theta^{\wedge} = 2/n(n\text{-}1)\ \Sigma_{1}\leq_{i}<_{j}\leq_{n} I_{ij}$ is unbiased [21]. Further, we are also able to evaluate the standard error of the sample density in this case in closed form to obtain $SE_{ustat}\ (\theta^{\wedge}) = 0.059$ [21].

If we treat the I_{ij}'s as if they were independent, we can apply conventional statistical methods such as the Binomial model to evaluate the standard error to obtain $SE_{Binam}(\theta^{\wedge}) = 0.0025$ [4]. The traditional methods clearly underestimates the variability of $\theta^{\wedge}$, yielding a bias almost 20 times bigger than the true variance.

In recent years, many methods and software packages have been developed to model social network features such as the social network density. To see how these methods work in this particular example, we applied some of the methods to the current context of social network density. One popular approach is the exponential random graph model employing Markov Chain Monte Carlo Maximum Likelihood estimation [23,24] We applied this approach as implemented in STATNET version 2014.2.0 Goodreau et al. [25] and obtained a standard error $SE_{Binam}(\theta^{\wedge}) = 0.003$. This standard error is just as biased as the one from the Binomial model.

To correct such downward bias, many packages offer Bootstrapped standard errors. In conventional statistical analysis, the Bootstrap is quite an effective approach to estimating standard errors by resampling observations Efron and Tibshirani [26], However, this popular approach does not apply to the current context, because it is premised upon the traditional paradigm of modeling within-subject attributes.

To confirm it, we applied this approach as implemented in the UCINET version 6.365 Borgatti, et al. [27] to the simulated data and obtained a standard error $SE_{Binam}(\theta^{\wedge}) = 0.041$ based on a Bootstrap sample of 5,000. The use of a large Bootstrap sample such as 5,000 in the current context is to reduce sampling variability and minimize its impact on estimated standard errors. Although much improved, the Bootstrap standard errors still show over 30% downward bias.

Discussion

Social network data offer a unique lens to investigating human interaction and its dynamic impact on individual outcomes. There is a burgeoning literature on both methods for modeling and applications of such methods to social network relationships and dynamics. Most of the methods provide descriptive statistics, as they only describe observed patterns among subject interactions (between-subject attributes) and relationships of such interactions to individual outcomes (within-subject attributes).A few notable exceptions include the exponential random graph model [23,28] and the network influence model [29,30]. As we discussed and illustrated using a simulation study, the exponential random graph model is fundamentally flawed, since it still applies conventional models for within-subject attributes to model between-subject attributes arising in social networks. The network influence model shares the same fate, since it too fails to correctly address the dependence among the between-subject attributes when modeling relationships between the between- and within-subject attributes in the social network context.

We proposed a new approach to address the dependence issue and illustrated it with the network density, a popular metric of social network connections. The new analytic paradigm is premised upon the theory of U-statistics, which is uniquely positioned to address interlocking relationships among between-subject attributes such as outcomes of connection between subjects within the context of social network density. The software for estimating social network density and simulating data for the simulation study discussed in this manuscript is available upon request.

In this paper, we focused on the social network density. In many studies involving social network data, interest lies in how individual outcomes (within-subject attributes) are influenced by their interactions with others in the network (between-subject attributes) and vice versa. For example, in studying disease contagion such as flu and depression, one may be interested in whether physical contact or friendship with a person with the disease increases the chance of contracting the disease. In this case, we may model disease infection as a function of connection (physical contact or friendship) using a regression with a within-subject attribute (disease infection) as a dependent and a between-subject attribute as an independent variable. This and other more complex models cannot be handled by the limited results in the theory of U-statistics. More general models extending the U-statistics such as the functional response models may be applied Kowalski & Yu et al. [21,22]. Research is currently underway to develop new statistical models for complex relationships involving interactions between the between- and within-subject attributes.

Acknowledgment

We are very grateful to Dr. Vincent Silenzio and Dr. Christopher Homan for their constructive comments about random graphs. We wish to thank to Dr. Eric Caine and the “Scholarly Pens” writing group, both of the Department of Psychiatry, for their constructive feedback as readers. The research was supported in part by grants DA027521 and GM108337 from the National Institutes of Health and by the University of Rochester CTSA award number UL1 TR000042 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Pentland AS (2013) Data driven society. Sci Am 309(4): 78-83.
2. Wasserman S, Faust L (1994) Social Network Analysis: Methods and Applications. Cambridge University Press pp. 825.
3. Cohen, R, Havlin S (2010) Complex Network: Structure, Robustness and Function. Cambridge University Press.
4. Tang W, Hua He, Xin M Tu (2012) Applied Categorical and Count Data Analysis. Chapman & Hall/CRC, Texts in Statistical science, pp. 384.
5. Feinberg ME, Riggs NR, Greenberg MT (2005) Social networks and community prevention coalitions. J Prim Prev 26(4): 279-298.
6. Valente TW, Chou CP, Pentz MA (2007) Community coalitions as a system: Effects of network change on adoption of evidence-based substance abuse prevention. Am J Public Health 97(5): 880-886.
7. Palinkas LA, Holloway IW, Rice E, Fuentes D, Wu Q, et al. (2011) Social networks and implementation of evidence based practices in public youth-serving systems: a mixed-methods study. Implement Sci 6(113).
8. Centola D (2010) The spread of behavior in an online social network experiment. Science 329(5996): pp 1194-1197.
9. Centola D (2011) An experimental study of homophily in the adoption of health behavior. Science 334(6060): pp 1269-1272.
10. Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: User movement in location-based social networks. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) Pp1082-1090.
11. Sadilek A, Kautz H, Silenzio V (2012) Modeling spread of disease from social interactions. In Sixth AAAI International Conference on Weblogs and Social Media (ICWSM).
12. El Sayed AM, Scarborough P, Seemann L, Galea S (2012) Social network analysis and agent based modeling in social epidemiology. Epidemiol Perspect Innov 9: 1.
13. Lakon CM, Hipp JR, Timberlake DS (2010) The social context of adolescent smoking: A systems perspective. Am J Public Health 100(7): 1218-1228.
14. Ramirez Ortiz G, Caballero Hoyos R, Ramirez Lez G, Valente TW (2012) The effects of social networks on tobacco use among high-school adolescents in Mexico. Salud Public a Mex 54(4): 433-441.
15. Luke DA, Harris JK, Shelton S, Allen P, Carothers BJ, et al. (2010) Systems analysis of collaboration in 5 national tobacco control networks. Am J Public Health 100(7): 1290-1297.
16. Archer J, Bower P, Gilbody S, Lovell K, Richards D, et al. (2012) Collaborative care for depression and anxiety problems. Cochrane Database Syst Rev 17: 10.
17. Kellam SG (2012) Developing and maintaining partnerships as the foundation of implementation and implementation science: reflections over a half century. Adm Policy Ment Health 39(4): 317-320.
18. Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19(3): 293-325.
19. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1(6): 80-83.
20. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics 18(1): 50-60.
21. Kowalski J, Tu XM (2007) Modern Applied U Statistics. Wiley New York 378.
22. Yu Q, Tang W, Kowalski J, Tu XM (2011) Multivariate U-Statistics:

A tutorial with applications. Wiley Interdisciplinary Reviews – Computational Statistics 3(5): 457-471.

23. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. Psychometrika 61(3): 401-425.

24. Snijders T A B (2002) Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3(2): 1-40.

25. Goodreau S M, Handcock M S, Hunter D R, Butts C T, Morris M, (2008) A STATNET Tutorial. J Stat Softw 24(9): 1-27.

26. Efron B, Tibshirani R J (1993) An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

27. Borgatti SP, Everett MG, Freeman LC (2002). UCINET for Windows: Software for social network analysis. Harvard MA: Analytic Technologies.

28. Strauss D, Ikeda M, (1990) Pseudo likelihood estimation for social networks. J Am Stat Assoc 85 (409), 204-212.

29. Christakis NA, Fowler JH (2008) The collected dynamics of smoking in a large social network. N E J Med 358: 2249-2258.

30. Cohen-Cole E, Fletcher JM (2008) Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. BMJ 337: a2533.

How to cite this article: Chen T, Lu N, White A, He H, Wu P. et.al. Social Network: A New Paradigm for Modeling Human Interaction: Implications for Statistical Inferences. Biostat Biometrics Open Acc J. 2017;1(1): 555551. DOI: 10.19080/ADOH.2016.01.555551