Sometimes Less Is More: When Aggregating Networks Masks Effects

Jennifer M. Larson[1(✉)] and Pedro L. Rodríguez[2]

[1] Vanderbilt University, Nashville, TN 37240, USA
jennifer.larson@vanderbilt.edu
[2] Meta, Menlo Park, CA 94025, USA

Abstract. A large body of research aims to detect the spread of something through a social network. This research often entails measuring multiple kinds of relationships among a group of people and then aggregating them into a single social network to use for analysis. The aggregation is typically done by taking a union of the various tie types. Although this has intuitive appeal, we show that in many realistic cases, this approach adds sufficient error to mask true network effects. We show that this can be the case, and demonstrate that the problem depends on: (1) whether the effect diffuses generically or in a tie-specific way, and (2) the extent of overlap between the measured network ties. Aggregating ties when diffusion is tie-specific and overlap is low will negatively bias and potentially mask network effects that are in fact present.

Keywords: Diffusion · Social networks · Tie aggregation · Multiplexity · Bias

1 Introduction

A large body of research aims to detect the spread of something through a social network. Be it voting behavior [18], news [12], new technology [8], disease [5], or patronage benefits [7], social networks are thought to matter because their ties—the relationships connecting pairs of people—serve as conduits that can spread ideas, information, goods, germs, and social judgment.

When researchers conduct an empirical test of this kind of spread, they collect a measure of the relevant social network [13]. Because "social relationship" is nebulous, the standard approach is to measure a few concrete types of social interactions. For instance, [8] use surveys to inquire about respondents' friends, family, potential money lenders, and potential problem solvers. Jennifer et al. [12] ask respondents about seven types of interactions, including visits to others' homesteads and shared meals. Banerjee [4] as about twelve, including borrowing rice and kerosene.

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 214–223, 2023.
https://doi.org/10.1007/978-3-031-21127-0_18

With multiple measures of social ties in hand, researchers typically take the union of these ties to construct one social network to be used for their analyses [3,4,8,11,12]. A tie in this aggregate network indicates that one or both of the tied individuals (depending on the coding rule the authors employ) reported having at least one of the measured relationships with the other. The logic that drives researchers to aggregate all measured ties is straightforward: each additional type of interaction contains weakly more information about the social relationship between two people, so the more interactions included, the better the measure of the social network.

However, we argue that this logic does not necessarily hold for all instances of something spreading through a network ("diffusion"). Some things may indeed diffuse generically, in which case one type of tie is as good as another for their spread. On the other hand, things may diffuse specifically along only certain kinds of ties. Take, for instance, the diffusion of useful, factual information, such as that a university is likely to declare a snow day tomorrow. We may expect this information to spread equally well along a tie that indicates shared membership in the rugby club and along a tie that indicates a twelve-year friendship. Now consider the diffusion of more sensitive information, such as that a student is considering reporting a professor for misconduct. In the latter case, students may opt to use their more trusted ties exclusively; one type of social tie would not be as good as another.

In short, ties that indicate a different kind of interaction between two people may also *work* differently to spread something, depending on the context. When diffusion is tie-specific, aggregating ties effectively adds measurement error to the network which, in an analogue to the regression context, can attenuate and mask true effects.

We begin by showing that different measures of social interactions among the same group of people can in fact pick up quite distinct views of the social network, using data from [12]. Then we demonstrate the problem of aggregating ties using a set of Erdős-Rényi random networks that vary in key ways. We show that when diffusion is specific, aggregating networks substantially attenuates the estimate of true effects. We further show that whether aggregation problems will arise depends on two factors: the amount of new information that that a candidate new tie type would add—a property we call "network overlap"—and the relative sizes of the networks being aggregated.[1]

Our results highlight the importance of careful theory and knowledge of context to underpin a network study. Determining whether the diffusion at hand is likely to be specific or generic relies heavily on theory and a deep qualitative understanding of local context. Without a guiding understanding of how something should, in principle, pass from person to person, researchers risk failing to detect it empirically.[2]

[1] This is distinct from the notion of overlap that refers to the extent to which two nodes share the same neighbors (see [15,17]).

[2] For a case in point, see [14], which shows that evidence of behavior spreading through the network is masked when the seven types of ties are aggregated. When disaggre-

2 Empirical Studies of Social Network Diffusion

Scholars are interested in "network effects" when they believe ties in a network can transmit something. In principle, theory should dictate precisely which ties are of interest. When theory is strong and the tie of theoretical interest has one clear operationalization, researchers are able to measure only a single type of tie. For instance, to study peer influence in adopting insurance among rice farmers, [6] measure one link type: household heads' close friends with whom they most frequently discuss rice production or other financial issues. To examine the advantage conferred to politicians by familial ties, [7] gather data exclusively on marriage connections between families in the Philippines.

Measuring a single tie type is the exception rather than the rule. When theory is not strong enough to identify one precise tie type that should be interest, or when a single type of tie in concept does not have a single obvious operationalization, researchers collect data on multiple types of ties.

For instance, in their study of farmers' decisions about sunflower crops in Mozambique, [3] aim to detect a relationship between one's peers' decisions and one's own by recording respondents' family members, friends, and neighbors. To study whether peers affect one's choice to give deworming medication to children in Kenya, [11] ask respondents to name the five friends and the five relatives they speak to most frequently, other social contacts whose children attend the local schools, and people with whom they speak about child health matters.

To learn about how social information spreads through Indonesian hamlets, [1] record both blood relatives and shared membership in social organizations. Jennifer et al. [12] measure the channels through which news may spread from person to person by asking Ugandan villagers about people with whom they spend time, share meals, exchange household visits, discuss religion, discuss politics, share secrets, and speak on the phone. Banerjee [4] measure twelve types of ties that may be responsible for spreading the word about a microfinance program: people who visit the respondent's home, people the respondent visits, kin, non-relatives with whom the respondent socializes, sources of borrowed money, sources of borrowed material goods, potential recipients of lent money, potential recipients of lent material goods, sources of medical advice, sources of general advice, receivers of advice, and prayer partners. Ferrali et al. [8] operationalize the ties that may spread information about new technology with friendship, family, potential money lenders, and potential solvers of problems about public services.

In all of the above cases of measuring multiple types of ties, the researchers take the union of the different types of ties to construct a measure of the "social network."[3] Doing so is the standard approach. At first blush, aggregating ties in this way seems justified since the resulting network contains the data's maximum information about social ties.

gated, the authors find evidence consistent with specific diffusion along the most intimate types of ties.

[3] In rare instances, researchers instead or also look at the different networks separately; see [2,14].

Moreover, a large literature has made clear that social relationships are complex. A relationship between two people can manifest in many ways, featuring a variety of kind of interactions. This "multiplexity" is a key feature of human networks (see [9,16]). In this view, social ties are multi-faceted and so a person who is my friend may also be my coworker as well as the person with whom I would discuss politics. We might even say that this relationship is especially rich because it features these three dimensions.

The difficulty arises when we flip this logic and look for evidence of a social tie by measuring these dimensions separately. Of course if everyone in a group of interest who are friends are also coworkers and are also political discusison partners, then there is no issue. Measuring one is the same as measuring all and taking their union. However, as we show below, when we pick a single dimension of a relationship—just political discussion partners, say—the network formed by those ties alone can look quite different from a network formed by a different dimension—working together, for instance. Given that different dimensions can add different information, it is worth considering whether aggregating across the different dimensions is always best.

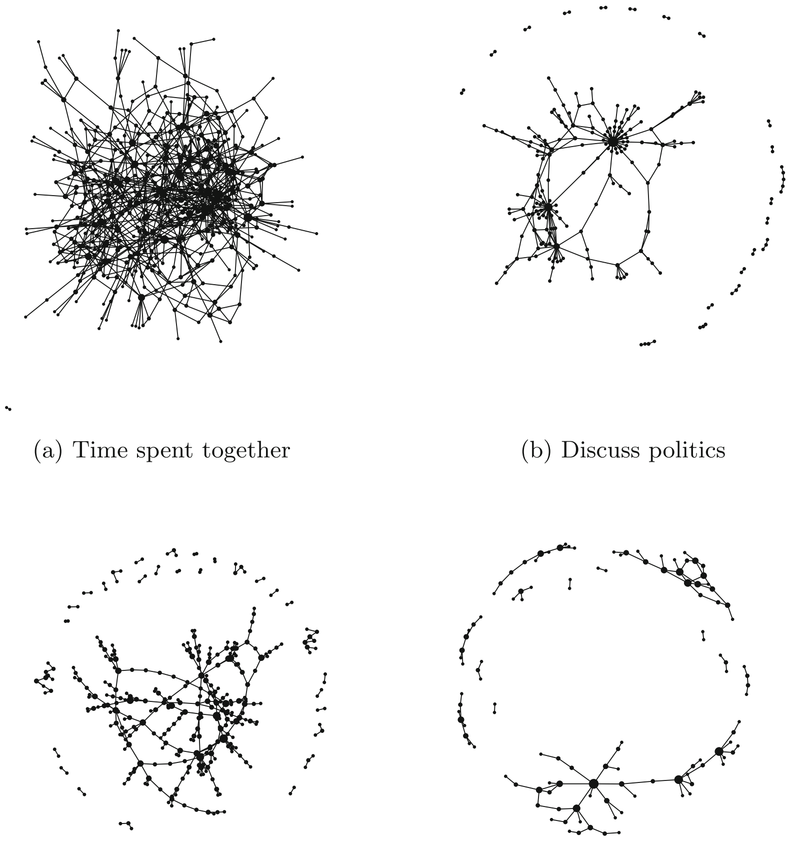
To preview our argument, consider the setting of [12] in which villagers can spread news to one another, including news that the researchers inserted into the network about an upcoming local event that would give out soap. Suppose that the true way that villagers pass on this kind of news is by telling anyone with whom they have any kind of social relationship. In such a case, measuring a variety of different social relationships and aggregating them would indeed maximize the relevant information.

Now suppose instead that villagers find events hosted by outsiders to be a political matter, worthy of discussion with one's politically-minded social ties and no others. Or perhaps villagers are concerned about their safety at an unusual event and will only discuss it with their elders. Different still, imagine that villagers fear that their participation, if discovered, would offend the local political elites, and so only spread word to their most trusted contacts. In each of these three scenarios, the true spread of information about the soap event only occurs along a specific kind of social tie—political discussion partners, elders, and the most trusted, respectively.

When this is the case, aggregating different kinds of ties can undermine the detection of network effects. In fact, by including the irrelevant information, the data may no longer reveal the true spread through the network.

3 Different Tie Types Generate Different Networks

To see that different measures of social ties can contain quite different information, consider the social network from one of the Ugandan villages in [12]. Seven different types of relationships are measured among the villagers by survey. Figure 1 highlights four of them. Each villager is asked to name other villagers with whom they regularly spend time (top left), discuss politics (top right), share secrets with (bottom left), and speak to on the phone (bottom right).



(a) Time spent together

(b) Discuss politics

(c) Share secrets

(d) Speak over the phone

Fig. 1. Different structural characteristics are apparent when social ties are measured with different kinds of interactions among the same people. Data from [12].

Although these questions are asked of the same villagers, each of the four dimensions of their social relationships looks very different when visualized as its own network. Taking their union to construct a single social network results in each dimension adding a different amount and type of information. If all of these relationships function the same for something spreading from villager to villager, then aggregation needs no further thought; they all provide a view of the same thing. But if diffusion in fact only occurs along ties based on a particular relationship (or even a strict subset of the relationships), then aggregation can mask our ability to detect it.

4 Theory

Suppose we are interested in evaluating whether peer effects play a role in diffusing a given piece of information in a rural village in Uganda. We hypothesize that once supplied to a random selection of "seed" villagers, the information will spread through each of these seed's contacts to the contacts of their contacts and so on. Contacts may vary in the type of tie that binds them. For example, contacts with whom an ego shares meals may differ from contacts with whom she works. In order to capture this heterogeneity suppose we measure two types of ties, ties of type A and ties of type B, the sets of which we denote S_A and S_B respectively. Some time after seeding the information we survey (non-seeded) villagers and inquire whether they have knowledge of the seeded information.

There is evidence of peer-effects if the likelihood of having knowledge of the information is positively correlated with the proportion of an ego's neighbors that were seeded. To evaluate this relationship we first need to determine which ties are relevant to quantifying the proportion of an ego's neighbors that were seeded. Ties of type A, of type B or both?

Suppose our seeded information did diffuse through the village's underlying social networks but only through ties of type A. In this case only the presence of a tie of type A between any two villagers in our sample is informative of peer effects. Taking the union of S_A and S_B is equivalent to introducing *noise* into our covariate measure. Given this equivalence we can appeal to the extensive literature on covariate measurement error to enumerate the consequences of aggregating networks in the presence of a tie-specific diffusion process. These are:

1. *Attenuation bias*: coefficient estimates are biased toward zero.
2. *Downward biased test statistics*: resulting in a higher probability of falsely failing to reject the null (type-II error).

The severity of these effects will depend on the true magnitude of the coefficient and the *noise-to-signal ratio*.[4] Unlike traditional covariate measurement error however—over which the researcher often has no control and little information—error resulting from mistakenly aggregating networks can be both avoided and quantified. Denote η as the noise-to-signal ratio and $n(\cdot)$ as the cardinality of a set. Returning to our hypothetical example we can show that:

$$\eta_{AB} = \frac{n(S_B - S_A)}{n(S_A)} \tag{1}$$

That is, the noise-to-signal ratio is equivalent to the number of ties in S_B not in S_A as a proportion of the number of ties in S_A. Notice, generally $\eta_{AB} \neq \eta_{BA}$, hence the subscript. It is useful to decompose η_{AB} into two components which we label size- and overlap-ratio. We define these as follows:

[4] For the derivation of this result for OLS see [10] and for logistic regression see [19].

Definition 1. Given two types of ties A and B, with respective sets S_A and S_B, define the ***size-ratio*** of S_A to S_B, as the ratio of the number of ties in S_B and number of ties in S_A. Formally:

$$s_{AB} = \frac{n(S_B)}{n(S_A)} \tag{2}$$

Definition 2. Given two types of ties A and B, with respective sets S_A and S_B, define the ***overlap-ratio*** of S_A to S_B as the proportion of ties in S_A also found in S_B. Formally:

$$o_{AB} = \frac{n(S_A \cap S_B)}{n(S_A)} \tag{3}$$

As with η_{AB}, neither ratio is symmetric with respect to its arguments hence the subscripts. It can be shown that η_{AB} is a function of these two ratios. Specifically,

$$\eta_{AB} = \frac{n(S_B - S_A)}{n(S_A)} = \frac{n(S_B) - n(S_A \cap S_B)}{n(S_A)} = \underbrace{\frac{n(S_B)}{n(S_A)}}_{\text{size-ratio}} - \overbrace{\frac{n(S_A \cap S_B)}{n(S_A)}}^{\text{overlap-ratio}} = s_{AB} - o_{AB} \tag{4}$$

All else equal, the larger the size-ratio, the more noise we are adding by taking the union. However, some of the ties in S_B may also be found in S_A. These ties do not add any noise.

5 Simulated Networks

In this section we illustrate the network aggregation problem using simulated random networks. Specifically we generate hypothetical networks with two types of ties, type A and type B, with varying size- and overlap-ratios. For a given population of nodes of size N,

1. Generate a list L of all potential ties between pairs of nodes in N.
2. Randomly select a subset S from L.
3. From S, randomly select a subset S_A, to make up the set of links of type A.
4. Remaining links in S along with o percent of links in S_A make up S_B.

We follow these steps to generate random networks for each combination of size-ratio s and overlap-ratio o such that $s \in \{1, 2, 3, 4\}$ and $o \in \{0, 0.25, 0.5, 1\}$.

For each of our generated networks we then select a random subset of nodes to serve as seeds, 5% in each. We then simulate a one-period diffusion process in which we assume information can only diffuse through ties of type A (in other words, we assume that diffusion is tie-specific). Knowledge of the information by i, Y_i, is a Bernoulli random variable taking value 1 with probability:

$$Pr(y_i = 1) = logit^{-1}(\beta_0 + X_i^A \beta_1)$$

where X_i^A is the proportion of i's neighbors that are seeds according to ties of type A. We set $\beta_0 = -1$ and $\beta_0 = 0.2$ and simulate 500 one-period diffusions for each size/overlap-ratio combination.[5] After each simulation we estimate the "noisy model":

$$Pr(y_i = 1) = logit^{-1}(\beta_0 + X_i^{agg}\beta_1)$$

where X_i^{agg} is the proportion of neighbors that are seeds according to the union of S_A and S_B. In practice, adding noise to the covariate measure reduces its variance, thereby reducing its explanatory power. Table 1 provides evidence for this intuition. Each value corresponds to the ratio of the average standard deviation of X_i^{agg} and X_i^A specific to a size- and overlap-ratio combination. As the size-ratio (overlap-ratio) of S_A to S_B increases (decreases) the smaller the variance of X_i^{agg} relative to the variance of X_i^A.

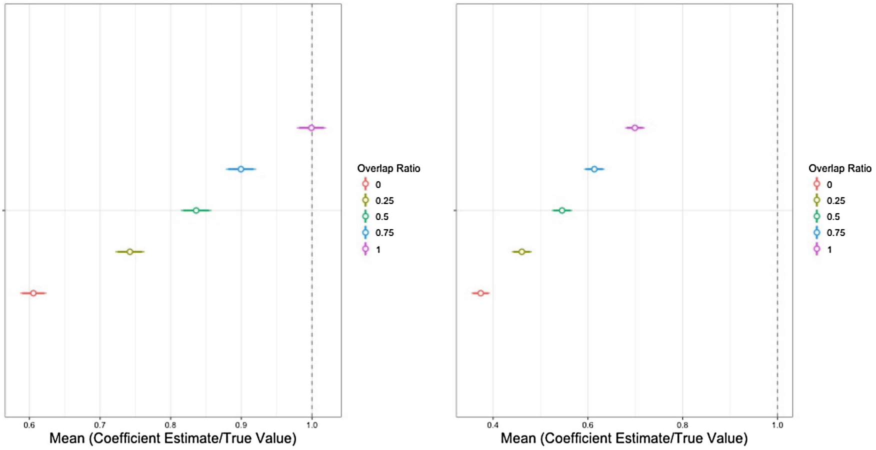
Table 1. Ratio of the mean standard deviations of X_i^{agg} and X_i^A.

Overlap-ratio						
Size ratio		0	0.25	0.5	0.75	1
	1	0.63	0.73	0.82	0.91	1
	2	0.50	0.55	0.60	0.64	0.68
	3	0.43	0.46	0.49	0.52	0.54
	4	0.38	0.40	0.42	0.44	0.46

Now define the *bias ratio* as the ratio of the estimated $\hat{\beta}_1$ and the true β_1. An unbiased estimator applied to the correct model yields on average a bias-ratio of 1. Figure 2 plots the distribution of the bias ratio for varying levels of overlap-ratio, holding the size-ratio equal to 1 and 4 in the left and right sub-figure respectively. Both figures show how the estimated effect is increasingly attenuated as the overlap-ratio between the two networks decreases. Moreover, even in the presence of perfect overlap—formally $S_A \subseteq S_B$—, we still observe attenuation as the number of uninformative ties increases in proportion to informative ties (see Fig. 2b).

As noted above, adding noise to the covariate measure also results in downward biased test statistics thereby increasing the probability of making type-II inferential errors. For each set of simulations -500 for each size-/overlap-ratio combination- we computed the proportion of regressions for which β_1 is statistically significant at 0.05. Table 2 summarizes these results. It is clear that as the size-ratio increases and overlap-ratio decreases the proportion of regressions for which we correctly reject the null decreases.

[5] The magnitude of β affects the severity of the attenuation bias, not whether it is present or not.



(a) Size-ratio = 1

(b) Size-ratio = 4

Fig. 2. Mean bias ratio as a function of size- and overlap-ratio.

Table 2. Proportion of simulations with significant β_1 (95% level)

Overlap-ratio						
Size ratio		0	0.25	0.5	0.75	1
	1	0.82	0.92	0.95	0.99	0.99
	2	0.58	0.73	0.86	0.92	0.96
	3	0.41	0.63	0.70	0.84	0.90
	4	0.35	0.53	0.67	0.74	0.86

6 Conclusion

Our article reexamines the standard approach of social network measurement, which defaults to taking the union of multiple types of social ties. This approach results in one aggregate "social network" that summarizes all information on social ties contained in the data. We point out that for many realistic cases, this logic is too simple, and can mask network effects that are in fact present.

We show that the decision of whether and how to aggregate ties should be informed by two factors: (1) whether the effect in question diffuses generically or in a tie-specific way, and (2) the extent of overlap between the measured networks. Aggregating ties when diffusion is tie-specific and overlap is low will negatively bias and potentially mask network effects that are in fact present.

Of course, knowing how the diffusion process should work is a substantial requirement. Our results highlight just how crucial strong theory and qualitative knowledge are for detecting meaningful network effects.

References

1. Alatas, V., Banerjee, A., Chandrasekhar, A.G., Hanna, R., Olken, B.A.: Network structure and the aggregation of information: theory and evidence from Indonesia. Am. Econ. Rev. **106**(7), 1663–1704 (2016)
2. Baldassarri, Delia: Cooperative networks: altruism, group solidarity, reciprocity, and sanctioning in Ugandan producer organizations. Am. J. Sociol. **121**(2), 355–395 (2015)
3. Bandiera, O., Rasul, I.: Social networks and technology adoption in northern Mozambique. Econ. J. **116**(514), 869–902 (2006)
4. Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O.: The diffusion of microfinance. Science **341**(6144), 1236498 (2013)
5. Bearman, P.S., Moody, J., Stovel, K.: Chains of affection: the structure of adolescent romantic and sexual networks. Am. J. Sociol. **110**(1), 44–91 (2004)
6. Cai, J., De Janvry, A., Sadoulet, E.: Social networks and the decision to insure. Am. Econ. J. Appl. Econ. **7**(2), 81–108 (2015)
7. Cruz, C., Labonne, J., Querubin, P.: Politician family networks and electoral outcomes: evidence from the Philippines. Am. Econ. Rev. **107**(10), 3006–37 (2017)
8. Ferrali, R., Grossman, G., Platas, M., Rodden, J.: Peer effects and externalities in technology adoption: evidence from community reporting in Uganda. SSRN (2018)
9. Gondal, N.: Multiplexity as a lens to investigate the cultural meanings of interpersonal ties. Social Netw. **68**, 209–217 (2022)
10. Greene, W.H.: Econometric Analysis. Pearson Education India (2003)
11. Kremer, M., Miguel, E.: The illusion of sustainability. Quarter. J. Econ. **122**(3), 1007–1065 (2007)
12. Larson, J.M., Lewis, J.I:. Ethnic networks. Am. J. Polit. Sci. **61**(2), 350–364 (2017)
13. Larson, J.M., Lewis, J.I.: Measuring networks in the field. Polit. Sci. Res. Methods **8**(1), 123–135 (2020)
14. Larson, J.M., Lewis, J.I., Rodríguez, P.: From chatter to action: how social networks inform and motivate in rural Uganda. Brit. J. Polit. Sci. (2021). https://doi.org/10.1017/S0007123421000454
15. Mattie, H., Onnela, J.-P.: Edge overlap in weighted and directed social networks. Netw. Sci. **9**(2), 179–193 (2021)
16. Mesch, G., Talmud, I.: The quality of online and offline relationships: the role of multiplexity and duration of social relationships. Inf. Soc. **22**(3), 137–148 (2006)
17. Peng, J., Agarwal, A., Hosanagar, K., Iyengar, R.: Network overlap and content sharing on social media platforms. J. Market. Res. **55**(4), 571–585 (2018)
18. Sinclair, B., McConnell, M., Michelson, M.R.: Local canvassing: the efficacy of grassroots voter mobilization. Polit. Commun. **30**(1), 42–57 (2013)
19. Stefanski, L.A., Carroll, R.J.: Covariate measurement error in logistic regression. Ann. Stat. 1335–1351 (1985)