Larson and Rodriguez
Applied Network Science (2023) 8:21
https://doi.org/10.1007/s41109-023-00546-7

Applied Network Science

RESEARCH Open Access

The risk of aggregating networks when diffusion is tie-specific

Jennifer M. Larson[1*] and Pedro L. Rodriguez[2]

*Correspondence:
jennifer.larson@vanderbilt.edu

[1] Department of Political Science, Vanderbilt University, Nashville, USA
[2] Core Data Science Meta, Menlo Park, USA

Abstract

Empirical studies of the spread of something through social networks, a process often called diffusion, tend to rely on network data assembled from the measurement of multiple kinds of social ties. These can be different kinds of relationships, such as friendship and kinship, or different instances of concrete interactions, such as borrowing money and eating meals together. Aggregating multiple measures of ties into a single social network has become standard practice, 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 bias and mask true network effects. We further demonstrate that the problem depends on: (1) whether the diffusion occurs generically or in a tie-specific way, and (2) the extent of overlap between the measured network ties. Aggregating multiple measures of ties when diffusion is tie-specific and overlap is low will, on average, attenuate and potentially mask network effects that are in fact present.

Keywords: Diffusion, Social networks, Peer effects, Tie aggregation, Multiplexity, Multilayer networks, Measurement error

Introduction

People are interconnected in social networks that can be comprised of a rich variety of relationships and facilitate countless types of interactions. Two people may be connected because they are friends, coworkers, blood relatives, members of the same organization, subscribers of the same newsletter, or possibly all of the above. Any one of these relationships may facilitate interactions that could range from enjoying free time together to sharing news of local events with one another to jointly plotting a coup. Studying what social networks do and when has become a vibrant literature that spans multiple disciplines (Bramoullé et al. 2016; Light and Moody 2020; Victor et al. 2017).

A large body of research in this tradition aims to understand whether and how the relationships in a social network allow something to spread from person to person, a phenomenon sometimes called "diffusion" (Burt 1980; Coleman et al. 1957; Valente 1996). Different studies focus on the spread of different things: voting behavior (Sinclair et al. 2013), news (Larson and Lewis 2017), new technology Ferrali et al. (2018), disease (Bearman et al. 2004), and patronage benefits (Cruz et al. 2017) are just a few examples. In each of these studies, 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 aim to empirically detect this kind of spread through real social networks, they collect a measure of the social network through which diffusion could occur. Of course, since there are many different relationships and interactions that could interconnect people (real networks are multilayered Kivelä et al. 2014), researchers have to make choices about which relationships to measure and how exactly to do so (Larson and Lewis 2020). The relationships that enter the data are ideally those that could in fact serve as channels of spread. The standard approach has become to measure a few concrete types of social interactions and infer the presence of the right relationship(s) from them. For instance, the approach in Ferrali et al. (2018) uses surveys to inquire about four different types of relationships and interactions: respondents' friends, family, potential money lenders, and potential problem solvers. The surveys used to gather network data in Larson and Lewis (2017) ask respondents about seven types of interactions, including visits to others' homesteads and sharing meals. In Banerjee et al. (2013), surveys ask about twelve different interactions, including borrowing goods such as rice and kerosene.

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 (Bandiera and Rasul 2006; Kremer and Miguel 2007; Banerjee et al. 2013; Larson and Lewis 2017; Ferrali et al. 2018). A tie in this aggregate network indicates that one or both of the tied individuals (depending on the coding rule the researcher employs) reported having at least one of the measured interactions with the other. The logic that drives researchers to aggregate all measured ties is straightforward: each 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. After all, social relationships tend to be complex and multi-faceted (Gondal 2022), so the more dimensions used to capture a possibly "multiplex" relationship, the better.

However, we argue that this logic does not necessarily hold for all instances of something spreading through a network.[1] Instead, the logic holds only for a particular case of diffusion that we call "generic." In such a case, something may spread just as easily along one kind of tie as along another; or along a tie whose presence was inferred from one type of interaction as opposed to another. On the other hand, it is possible for certain things to spread only along certain kinds of ties; we refer to this as the case of "specific diffusion."

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, a tie that indicates currently being in the same math class together, and a tie that indicates a twelve-year friendship. This simple piece of news may be passed just as freely along any

[1] We are not the first to raise the issue that aggregation can be misleading; see also De Domenico et al. (2015); Kivelä et al. (2014); Cozzo et al. (2013)

of these social connections. Contrast that scenario with the diffusion of more sensitive information, such as that a student is considering reporting a professor for misconduct. In the latter case, a student may not feel so free to spread the word to just anyone they know socially. They may opt to exclusively use their more trusted ties to share, perhaps selecting only those with whom they share a long friendship. In such a case, one type of social tie is not as good as another for diffusion.

In short, ties that indicate a different kind of interaction between two people may also *work* differently to spread something, depending on the context (Aral and Van Alstyne 2011; Granovetter 1973; Larson 2017). The problem is that when diffusion is in fact tie-specific, aggregating different kinds of ties effectively adds measurement error to the network which, in an analogue to the regression context, can attenuate and mask true effects.

Our approach is to build on Larson and Rodríguez (2022) to characterize this problem theoretically, demonstrate it hypothetically, and then confirm it using real network data for 75 different social networks that each contain 12 different measures of ties. We begin by noting that different measures of social interactions among the same group of people can in fact pick up quite distinct views of a social network, using data from Larson and Lewis (2017). Then we characterize the problem theoretically, identifying two features of networks that affect the extent of the downside to aggregating them: the amount of new information that a candidate new tie type would add—a property we call "network overlap"—and the relative sizes of the networks being aggregated—a property we call "size ratio".[2] Next, we demonstrate the problem of aggregating ties using a set of null networks generated as Erdős-Rényi random networks with particular constraints on link formation to ensure that the resulting networks vary in the two features our theory identifies. This exercise shows that when diffusion is specific, aggregating networks will, on average, substantially attenuate the estimate of true effects. Finally, we validate the claim that this problem could arise in networks that are more realistic depictions of settings in which diffusion might occur by repeating the exercise on 75 real social networks.

Our results caution against defaulting to aggregating measures of different ties. To avoid attenuation bias, researchers need to first consider the type of diffusion that may be present. These findings highlight the importance of careful theory and contextual knowledge in grounding an empirical 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, and of how the networks under consideration function to facilitate this, researchers risk estimating biased effects—downward biased in most cases, though we show there are caveats to this—or failing to detect an effect altogether.[3]

[2] Note that this is distinct from the notion of "overlap" that refers to the extent to which two nodes share the same neighbors (see Mattie and Onnela (2021); Peng et al. (2018)).

[3] For a case in point, see (Larson et al. 2021), which shows that evidence of behavior spreading through the network is masked when the seven types of ties are aggregated. When disaggregated, the authors find evidence consistent with specific diffusion along the most intimate types of ties.

Table 1 Examples of multiple tie types measured in empirical network studies

Context	# Ties	Tie types measured
Agriculture in Mozambique Bandiera and Rasul (2006)	3	Family, Friends, Neighbors
Technology in Uganda Ferrali et al. (2018)	4	Family, Friends, Lenders, Problem Solvers
Public Goods in Ghana Atwell and Nathan (2022)	4	Family, Friends, Lenders, Problem Solvers
News in Uganda Larson and Lewis (2017)	7	Share Meals, Share Secrets, Discuss Religion, Discuss Politics, Call on Phone, Visit, Spend Time
Program Uptake in India Banerjee et al. (2013)	12	Visit, Visited By, Kin, Socializes With, Borrow Money, Borrow Goods, Lend Money, Lend Goods, Medical Advice, General Advice, Give Advice, Prayer Partner

Empirical studies of social network diffusion

A pair of individuals can be interconnected by a variety of different types of relationships, and can interact in many different ways. Consequently, saying we want to study "the" social network among a group of people is an ambiguous guide for empirical research. Researchers have options for which relationships they wish to capture in their measure of a social network, and often choose to operationalize the relationships of interest with concrete activities or interactions; in fact using a survey to ask about these tangible tasks can improve the quality of network data (Larson and Lewis 2020).

In principle, theory should dictate precisely which relationships or interactions are of interest for any empirical study of diffusion. When theory is strong and the network of theoretical interest has one clear operationalization, researchers are able to measure a single type of relationship or interaction. For instance, to study the role that peer influence may play in rice farmers' decisions to adopt insurance, the authors in Cai et al. (2015) measure one tie type: household heads' close friends with whom they most frequently discuss rice production or other financial issues. This is the one set of interactions that indicate the channel through which insurance adoption behavior may spread according to their theory. To examine the advantage conferred to politicians by familial ties, (Cruz et al. 2017) gather data exclusively on marriage connections between families in the Philippines. Theory suggests that patronage benefits may flow from politicians to families through these intermarriages, and so these are the only relevant ties to include.

In practice, measuring a single type of relationship or interaction, which we can broadly refer to as a tie type, is the exception rather than the rule. Often, theory is not strong enough to dictate one precise tie type that should be of interest, suggesting merely that some kind of social interaction should matter. In other cases, the theory may be precise and yet have no one obvious operationalization. In cases like these, researchers collect data on multiple types of ties. Table 1 gives some examples.

In their study of farmers' decisions about sunflower crops in Mozambique, (Bandiera and Rasul 2006) 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, (Kremer and Miguel 2007) 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, (Alatas et al. 2016) record both blood or marriage relatives and shared membership in social organizations. The channels through which news may spread from person to person are measured in Larson and Lewis (2017) 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. Twelve types of ties that may be responsible for spreading the word about a microfinance program are measured in Banerjee et al. (2013): 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. The ties that may spread information about new technology in Uganda are operationalized in Ferrali et al. (2018) with friendship, family, potential money lenders, and potential solvers of problems about public services. The study of how networks work to facilitate the production of public goods in Atwell and Nathan (2022) uses the same four tie types.

In 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,"[4] an approach which has become standard.

Multilayer networks

The fact that multiple kinds of ties might all be relevant for the spread of something through a network is the subject of a growing literature focused on "multiplex" and "multilayer networks" (Bianconi 2018; Dickison et al. 2016; Boccaletti et al. 2014; Kivelä et al. 2014). In a multilayer network, each type of tie is included in the network object as a separate layer. This literature has developed tools to describe these kinds of networks and model processes on them. These tools treat the ties in the network as separate layers that may all be important, possibly in different ways.

Studies using these tools have established that accounting for multiple, separate layers can reveal important insights about network processes. For instance, models of diffusion on multilayer networks account for the possibility that more than one layer (tie type) may be responsible for spreading something (Salehi et al. 2015). They reveal that when more than one layer of ties helps to spread of something, diffusion can occur even more quickly (Gomez et al. 2013), and depending on the structure of each of the layers, congestion can be especially prevalent (Solé-Ribalta et al. 2016). Some models also account for the possibility that the spread in question could occur more easily within a layer than

[4] In rare instances, researchers instead or also look at the different networks separately, for instance in Atwell and Nathan (2022); Baldassarri (2015); Larson et al. (2021). The standard approach is to use the union of all measured tie types on the grounds that it contains maximal social information.

across layers (De Domenico et al. 2015), or could entail a cost or overhead when switching layers (Min and Goh 2014; Cozzo et al. 2013). In all of these models, the presumption is that the layers included are relevant to the process at play.

This work has revealed important insights about situations in which there are multiple, known layers through which diffusion can occur. Researchers facing a variety of possible tie types to include in a study are effectively facing the possibility of treating their network as multilayered. Whether the researcher plans to consider each layer separately or ultimately aggregate them, they are confronted with a key question: which layers should be included? The multilayer networks literature offers some guidance when the goal is to create the best representation of a structure, or highlight the presence of communities most clearly (Cardillo et al. 2013; De Domenico et al. 2015). However, we show that this question is especially poignant when the researcher wants to empirically detect diffusion by measuring a real social network and observing an outcome that may or may not have spread through it. The question that can make or break the ability to detect diffusion is: *which* layers are the relevant ones?

Thanks to the expansion of methods for collecting network data, researchers tend to face an abundance of options of layers that they could include in their network study. The layers are not only numerous; they are also highly variable in the kinds of relationships they capture and in how distinct any one is from another. Some directly measure different relationships: friends, coworkers, kin. Others measure activities and interactions that could but do not have to indicate different relationships: visiting homesteads, borrowing kerosene. Which ties ought to be included as layers in a network study if the goal is to observe an outcome and infer the presence of diffusion? This question warrants careful consideration because including the wrong ones can compromise the study.

The many perils of aggregation

It is well know that when a network is truly comprised of multiple layers that are each relevant to a spreading process, aggregation is common practice but not ideal. The problem is that each layer may contribute something different to the overall network structure which may be concealed when included with other layers that have different structural features (De Domenico et al. 2015; Kivelä et al. 2014; Cozzo et al. 2013). We argue that an additional problem with aggregation arises from a different source: the inclusion of irrelevant ties.

At first blush, aggregating ties by taking their union seems unproblematic since the resulting network contains the data's maximum information about the presence of social ties. After all, 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 kinds of interactions. This "multiplexity" is a key feature of human networks (Mesch and Talmud 2006; Gondal 2022). 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 all three of these 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 as well as political discussion partners, then there is no

issue. Measuring one is the same as measuring all and taking their union. The view that we would have of the overall network structure would be the same. 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 measuring a different dimension– working together, for instance. Given that different dimensions add different information, it is worth considering whether aggregating across the different dimensions is always best.

To preview our argument, consider the setting of Larson and Lewis (2017) 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 information relevant to diffusion.

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 would only occur along a specific kind of social tie; political discussion partners, elders, and the most trusted, respectively.

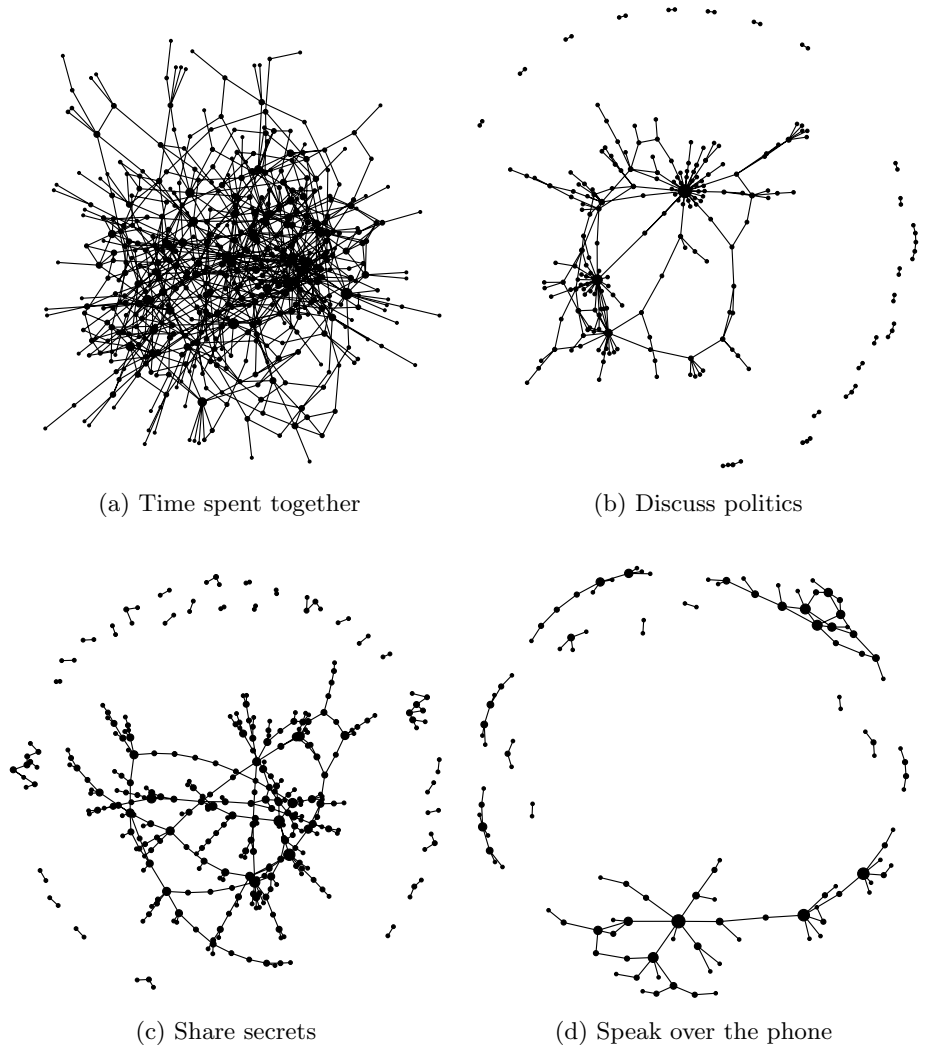
When this is the case, aggregating different kinds of ties can undermine the estimation and detection of network effects. Taking the union of all measured types of social ties does maximize information in a sense, but does so by including information irrelevant to the diffusion process. And as we show below, it can be the case that including the irrelevant information makes it impossible for the researcher to detect the true spread through the network.

Different tie types generate different networks

If every tie type measured among a group of people gave a roughly equivalent picture of the social network among them, then we would not need to worry about this issue of diffusion potentially being specific to certain kinds of ties. So we first show that real measures of different tie types lead to views of networks with substantially different structures.[5]

Take the tie types measured in one of the Ugandan villages in Larson and Lewis (2017). Seven different tie types are measured among the villagers via a name-generator survey. Figure 1 shows four of them for illustration. For these tie types, each villager was asked to name other villagers with whom they regularly spend time (top

[5] Others have made this point in other contexts as well. For instance, the networks comprised of the different tie types measured among players of a massive multiplayer online game also look quite different, especially the ties that are indicative of cooperative relations compared to those that are more negative (Szell et al. 2010). The networks of Twitter users who tweet about protests are different depending on the relation used to define the link (González-Bailón et al. 2011). The networks of countries look different when using trade links as opposed to alliances (Maoz 2012).



(a) Time spent together (b) Discuss politics

(c) Share secrets (d) Speak over the phone

Fig. 1 Same village's network measured with four different tie types

left), discuss politics (top right), share secrets (bottom left), and to whom they speak on the phone (bottom right).

Although these views all show the result of network questions asked of the same set of villagers, each of the four tie types produces a very different view of the village's network. Moreover, the difference in views is not merely one of the number of ties. Rather than one set of ties being a tidy subset of another, these views contain some of the same ties but far from all of them, and bring in many new ties, which results in different networks with quite different structural features.

Consequently, when the union of ties as different as these is taken to construct a single social network, each dimension is adding a different amount and type of information. If all of these relationships function the same for something spreading from villager to villager, then this issue is not important; all of the tie types would provide additional information relevant to the question at hand and so including them would add value. 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.

Theory

In this section we make the notions of generic and tie-specific diffusion precise and characterize features of the networks generated by each tie type that matter for whether aggregating them will mask diffusion or not.

Consider a set of nodes $N = \{1, \ldots, n\}$. Let S_K denote the set of ties of type K among nodes in N. We can imagine a situation in which we have multiple different types of ties measured. Let $\mathcal{K}$ be the set of all of the types measured, so that we have sets of ties $S_K, S_{K'}, \ldots$ with $K, K' \in \mathcal{K}$ that we could either treat as separate networks or aggregate. S_K could be shared meal ties, $S_{K'}$ visit each other ties, and so on.

When our interest is in diffusion (we could also call this an interest in "peer effects"), we can think about two extreme cases of the way the spread of something interacts with the different network types. The key distinguishing factor will be whether and how well spread can occur along each of the tie types. Let $Pr(i \rightarrow j)$ stand for the probability that something originating with node i spreads to node j, and $ij \in S_K$ indicate a tie between i and j in the set of ties of type K.

Definition 1 Generic Diffusion: A diffusion process is generic if $Pr(i \rightarrow j)$ for $ij \in S_K$ is equal to $Pr(i \rightarrow j)$ for $ij \in S_{K'}$ for all $K, K' \in \mathcal{K}$.

In other words, at one extreme, a diffusion process is generic if the type of tie has no bearing on the probability of something spreading between two nodes. However likely it is for the thing to spread from i to j if they were connected by a tie of type K, it would be equally likely to spread if they were connected by a tie of type K'.

At the other extreme, the diffusion process may depend on the tie type. The starkest version of this would be the situation in which diffusion could only occur along one of the types of ties.

Definition 2 Specific Diffusion: A diffusion process is specific to a tie type if there is a $K \in \mathcal{K}$ such that $Pr(i \rightarrow j) > 0$ for $ij \in S_K$ and $Pr(i \rightarrow j) = 0$ for $ij \in S_{K'}$ for any other $K' \neq K$.

A diffusion process is specific if spread can only occur through one type of tie.[6] Our claim is that when diffusion is specific, aggregating tie types can be problematic.

Exactly how problematic depends on the extent to which the irrelevant ties add different information to the network, and the volume of the different information. To be more specific, let's take a simple case where there are two types of ties in $\mathcal{K}$, and let's label them A and B to simplify notation. Again, these could be two different types of

[6] Of course these are the most extreme cases of what we might mean by generic and specific. We might prefer to call a diffusion process specific if there exists a strict subset of tie types such that types outside of this subset participate in spread with probability 0. The logic of what follows will hold if we use this as our definition; we would simply relabel the indicators for the relevant ties and the irrelevant ones. We could also imagine these two definitions as endpoints on a continuum and the diffusion process being relatively more or less specific depending on whether the probability of spread along the tie types is very different or similar across types. Demonstrating the problem is simpler with these stark definitions, but again, the logic continues to hold for the comparison of relatively specific to relatively generic diffusion.

relationships, say friendship and kinship, or two different types of interactions, say borrowing money and speaking on the phone. Now S_A is the set of all ties of type A and S_B the same for type B.

Imagine we want to measure the diffusion of something, say information, among our set of nodes N, say villagers. A standard design for this sort of study would be to experimentally seed information with a few villagers, wait some amount of time, conduct a survey to determine who received the information, and then measure the statistical relationship between the proportion of a node's ties who were seeded and whether that node received the information.

Suppose our diffusion is tie-specific, and without loss of generality, suppose it only spreads through ties of type A. In such a case only the presence of a tie of type A between any two nodes in our sample can contribute to the diffusion process. That is, only ties of this type are 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 establish the consequences of aggregating networks in the presence of a tie-specific diffusion process. These are:

(1) *Attenuation bias*: coefficient estimates are, on average, 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*.[7] 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 $\#(\cdot)$ as the cardinality of a set. Returning to our hypothetical example we can show that:

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

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:

Definition 3 Given two types of ties A and B, with respective sets S_A and S_B, define the ***size-ratio*** of S_B to S_A, as the ratio of the number of ties in S_B as a proportion of the number of ties in S_A. Formally:

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

Definition 4 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:

[7] For the derivation of this result for OLS see (Greene 2003) and for logistic regression see (Stefanski and Carroll 1985).

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

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,

$$\eta_{AB} = \frac{\#(S_B - S_A)}{\#(S_A)} = \frac{\#(S_B) - \#(S_A \cap S_B)}{\#(S_A)} \tag{4}$$

$$= \underbrace{\frac{\#(S_B)}{\#(S_A)}}_{\text{size-ratio}} - \underbrace{\frac{\#(S_A \cap S_B)}{\#(S_A)}}_{\text{overlap-ratio}} = s_{AB} - o_{AB} \tag{5}$$

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

Characterizing and verifying the aggregation problem

The last section identifies two network features that determine the extent of the problem with aggregating tie types in a situation of tie-specific diffusion. Next we demonstrate how these features relate to problems of inference by generating networks that vary in these two features and simulating attempts to draw inferences about the diffusion process. Specifically, our approach will do the following:

(1) Stipulate a known diffusion process
(2) Suppose diffusion only spreads along one type of tie ("tie-specific")
(3) Generate hypothetical networks with two types of ties that vary in the ways the last section identified as important
(4) Simulate specific diffusion along one tie type
(5) Aggregate the two tie types
(6) Test for diffusion in the aggregated network
(7) Record bias and type-II error

Generated networks

In this section we illustrate the network aggregation problem using hypothetical networks for our simulated diffusion process.

Continuing our running example, we will simulate the simple diffusion process that occurs after seeding information with a small, random selection of villagers. Ties will continue to have two types, A and B, and we will again suppose that the diffusion process is tie-specific, working only along ties of type A and not B. There is said to be evidence of peer-effects if the likelihood of having knowledge of the information is positively correlated with the proportion of an ego's network neighbors that were seeded with the information.

Specifically, we generate hypothetical networks with two types of ties, type A and type B, with varying size- and overlap-ratios, the two features that the last section identified

Table 2 Ratio of the mean standard deviations of X_i^{agg} and X_i^A

		Overlap-ratio				
		0	**0.25**	**0.5**	**0.75**	**1**
Size	1	0.63	0.73	0.82	0.91	1
Ratio	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

as relevant. We begin by defining a population of nodes N of size n and proceed with the following network generating process that will allow us to stipulate a size- and overlap-ratio:

(1) Generate a list L of all potential ties between 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 ties of type A.
(4) The remaining ties in S along with o percent of ties in S_A make up S_B.

We follow these steps to generate random networks for each combination of s and o such that $s \in \{1, 2, 3, 4\}$ and $o \in \{0, 0.25, 0.5, 1\}$. From equation 4 it can be shown that for a given combination of size- and overlap-ratio to characterize the relationship between S_A and S_B, the following equality must hold:

$$\frac{n(S_A)}{n(S)} = \frac{1}{1 + \eta_{AB}} = \frac{1}{1 + s_{AB} - o_{AB}} \tag{6}$$

For each of our generated networks we select a random subset of nodes, in both S_A and S_B, to serve as seeds, 5% in each. We then simulate a simple one-period diffusion process in which we assume that something– in our running example that thing is information– can only diffuse through ties of type A. 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 $\beta_1 \in [0.05, 0.4]$ in increments of 0.05, and simulate 100 one-period diffusion processes for each size/overlap-ratio and effect-size combination. After each simulation we then 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. This reflects researchers using the union of ties in their analyses in a world in which the diffusion was in fact tie-specific. In practice, adding noise to the covariate measure reduces its variance, thereby reducing its explanatory power. Table 2 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

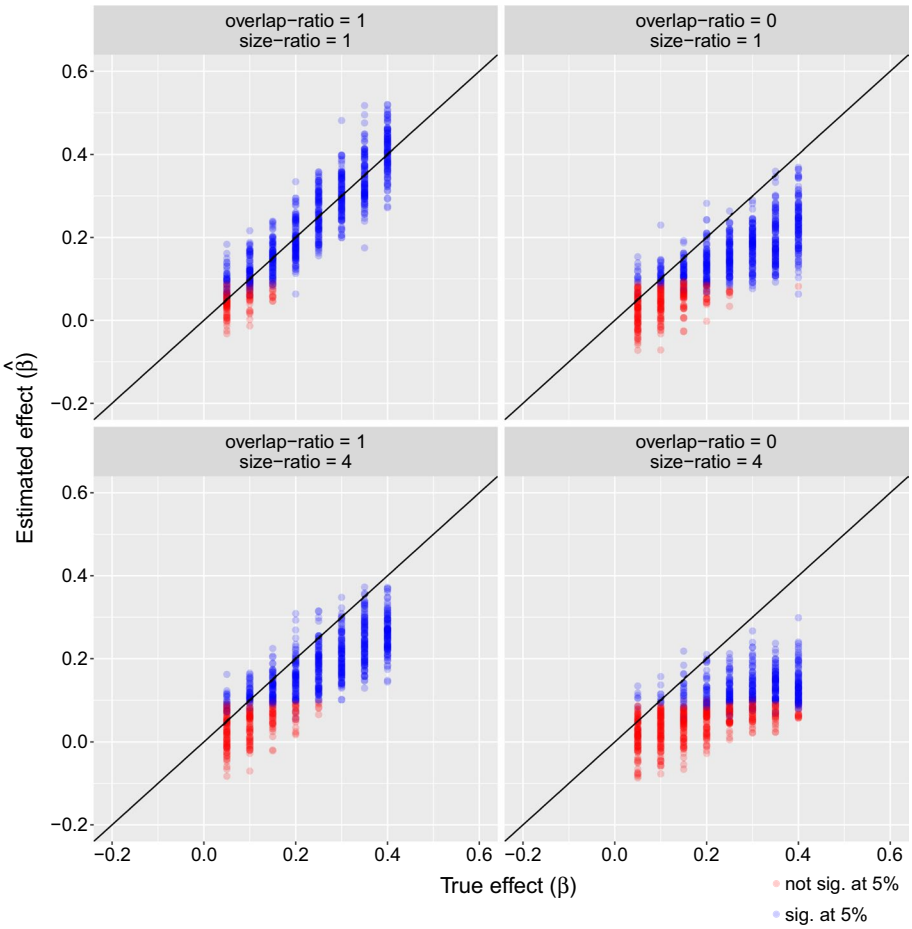


Fig. 2 Estimated versus true effects as a function of size- and overlap-ratio

(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.

Figure 2 plots the estimated effects against the true effects, with a 45° line to facilitate comparison. First, focusing on the top left panel we observe that on average when the overlap-ratio and size-ratio are both 1 –perfect overlap and equally sized networks–, aggregation is not an issue. On average, the estimated effect will be equal to the true effect. However, once we move away from perfect overlap and equally sized networks –right panels and lower panels respectively– we observe the estimated effect is on average biased downwards (below the 45° line).

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. We confirm this in Fig. 2, with the proportion of non-significant estimates –observations in red– increasing with lower overlap- and higher size-ratios.

While attenuation is, on average, more severe the larger the true effect size, the increase in the likelihood of type-II inferential errors is more pronounced the smaller the true effect size. Importantly, and contrary to conventional wisdom, in the presence of small true effect sizes and noisy covariates, significant estimates are more

likely to be magnifying rather than attenuating the true effect (see bottom right-hand panel). As such, we caution researchers against interpreting observed small significant effects in the presence of noisy covariates as lower bounds.[8]

Real networks

The last section illustrated the potential attenuation effect of network aggregation in simulated hypothetical networks that vary in the two network features that theory suggests matter: size- and overlap-ratio. Since those were artificial networks, that section leaves open the possibility that real networks do not have tie types so different that they have size- and overlap-ratios of sufficient magnitude to cause meaningful problems in inference if they were aggregated.

To address this concern, in this section we illustrate the network aggregation problem using real social networks measured in the field. We employ a dataset that includes social networks measured in 75 villages in India Banerjee et al. (2014), each with measures of twelve different types of ties. If we are to observe attenuation bias in real networks then it must be the case that there is significant variation in overlap- and size-ratios in real networks. Figure 3 plots these ratios for all possible tie type pairs within a village for all villages.[9] Significant variation can be observed for both ratios. Real measures of tie types generate a wide range of size- and overlap-ratios. The next question is whether these size- and overlap-ratios would be of the right magnitude to mask evidence of diffusion in the case where diffusion were tie-specific.

To illustrate the attenuation effect in this large set of real networks, we run the same one-period diffusion process as with our generated random networks. Specifically, for each village, we simulate 500 one-period diffusion processes, along each of the 12 tie types separately to again mimic tie-specific diffusion. We set $\beta_0 = -1$ and $\beta_1 \in [0.05, 0.4]$ –the min and max true effects evaluated in the our simulated networks.

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 4 displays the resulting average bias ratio as a function of the overlap-ratio and size-ratio quartiles for each village-tie type combination, split by true effect size. The dashed red line indicates a bias-ratio of 1. Consistent with theory and our results using simulated networks, we observe that a smaller (larger) overlap-ratio (size-ratio) is associated with more pronounced downward bias, indicated by a larger proportion of the observations lying to the left of 1. Moreover, while we observe attenuation for both small (left hand column) and large (right hand column) true effect sizes, attenuation is more severe for the latter. Indeed, as observed with simulated networks, in the case of small true effect sizes (and low power more generally) estimates can often lie above the true value.

Tables 3 and 4 show by size- (rows) and overlap-ratio (columns) quartiles, the proportion of regressions for which we correctly reject the null of no peer-effects, given a true effect size of 0.4 and 0.05 respectively. Again, while attenuation is more

[8] Loken and Gelman (2017) show how in the presence of low power and noisy covariates, statistically significant estimates are more likely to be magnifying rather than attenuating true effects. Their analysis focuses on low power as a result of sample size rather than small true effect size.

[9] Note, size- and overlap-ratios are not symmetric, as such for each tie-pair within a village we compute and plot ratios in both directions. For ease of reading, we exclude observations with size-ratio > 5. The maximum size-ratio is 43.54.

Table 3 Proportion of simulations with significant β_1 (95% level, $\beta_1 = 0.4$)

		Overlap-ratio quartile			
		1	**2**	**3**	**4**
Size	1	0.455	0.617	0.640	0.639
Ratio	2	0.402	0.537	0.548	0.578
Quartile	3	0.329	0.472	0.504	0.484
	4	0.279	0.352	0.393	0.370

Table 4 Proportion of simulations with significant β_1 (95% level, $\beta_1 = 0.05$)

		Overlap-ratio quartile			
		1	**2**	**3**	**4**
Size	1	0.049	0.052	0.057	0.055
Ratio	2	0.511	0.050	0.050	0.052
Quartile	3	0.043	0.048	0.050	0.047
	4	0.041	0.045	0.042	0.046

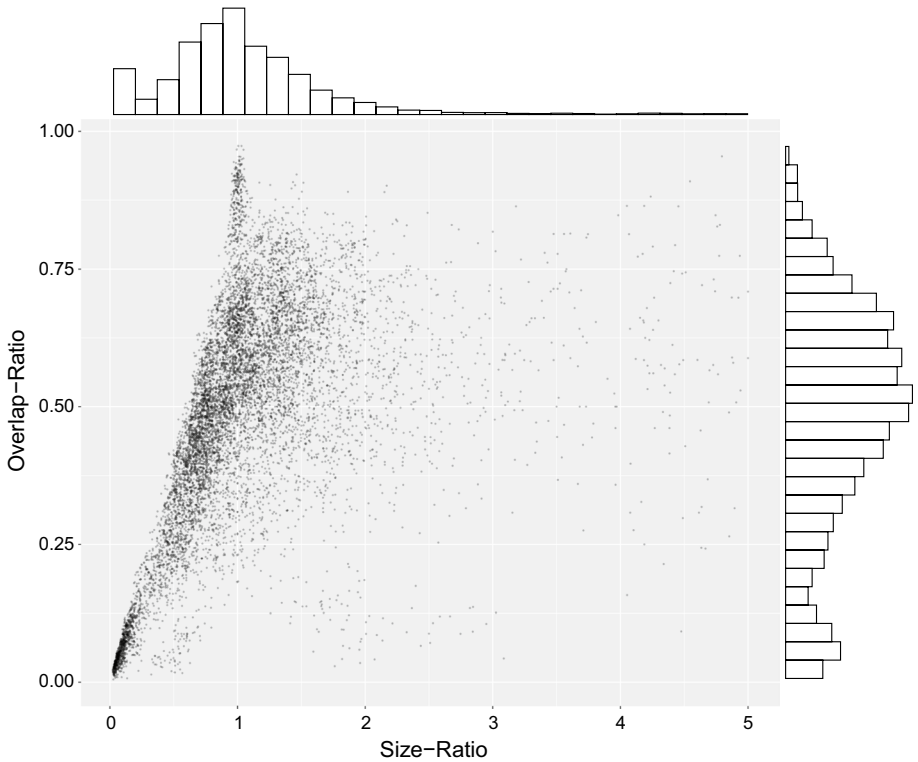


Fig. 3 Significant variation in overlap- and size-ratio in real social networks

significant in the case of larger true effect sizes, type-II inference errors are more likely given small true effect sizes. In both cases inference errors are more pronounced in the presence of low overlap-ratio and high size-ratio.

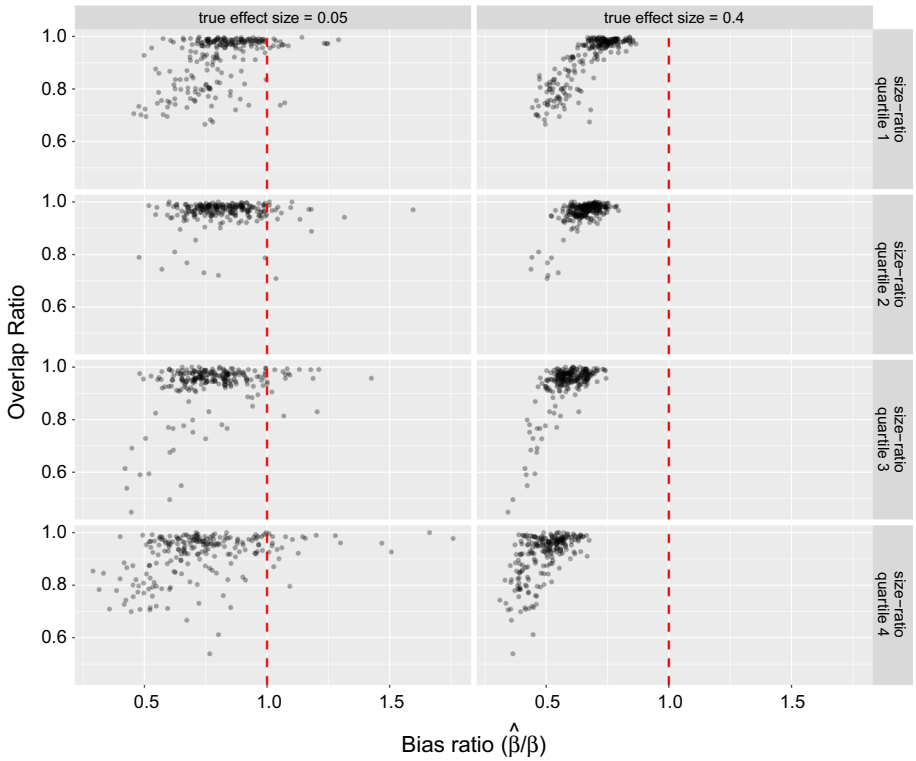


Fig. 4 Mean bias ratio as a function of overlap- and quartiles of size-ratio for real networks. The dashed red line indicates a bias-ratio of 1 (no bias on average)

Discussion

Significant variation in the size- and overlap-ratio of measured ties is usually a desirable result. Indeed the reason for measuring several types of ties is to capture as many of the potential paths as possible through which diffusion can occur. If all measured ties perfectly overlapped and brought the same number of ties to the data, then we would be gaining no new information by measuring different ties while still paying the cost of including extra items in our surveys.

However, the simulations on generated and real networks above flag two important considerations that should inform the decision to aggregate network ties: how should the diffusion in question spread through the network in theory, and how much additional information is contained in any measure of the network tie type. Table 5 summarizes these considerations.

In short, some diffusion processes can be specific to certain kinds of ties. There are contexts in which this claim is straightforward. If our interest were in the short-term spread of the flu, we would expect ties that indicate shared physical space to be relevant, and might expect ties that indicate shared long-distance phone calls to be irrelevant. However, in plenty of contexts it might not be as obvious when irrelevant ties are present in the data. And, to complicate matters further, although we present the distinction as discrete, there is likely in reality a continuum of relevance, and it is up to the researcher to know when ties are below some relevance threshold.

Table 5 Consequences of aggregating multiple network measures for detecting true network effects. When diffusion is specific, aggregating ties can mask effects; the consequences are worse when there is low overlap between the different tie types

	Low overlap	**High overlap**
Specific diffusion	Severely masks	Lightly masks
Generic diffusion	Fine	Fine

If researchers know much more about the process governing spread in their empirical context, then they can turn to models of multilayer networks that avoid the need to aggreagate the ties in the first place (Bianconi 2018; Dickison et al. 2016; Boccaletti et al. 2014; Kivelä et al. 2014). These models stipulate how each of the layers contributes to the spread on the network. For instance, if researchers know the processes is such that a node receives the thing spreading once a threshold is exceeded, and each layer can contribute independently to meeting the threshold, they can use a model like that in Brummitt et al. (2012). If the researcher knows the different tie types should be weighted differently in meeting a threshold and knows their weights, they can use (Yağan and Gligor 2012). Researchers can weight the ability of the thing to spread within a layer differently from the ability to cross layers by following (De Domenico et al. 2015). Researchers can also account for multiple things spreading at once, which opens the door to contrasting coupling (something in one layer responding to something different spreading an another layer) and switching (something jumping from one layer to another) (Brodka et al. 2020). When there are different characteristics of the tie types that could affect the flow, such as their sign (indicating relationships that are positive or negative), researchers can also account for these (Leskovec et al. 2010). In short, the more that is known theoretically about the true underlying process on the network, the better any empirical modeling effort will be.

The study of peer effects to date has remained relatively agnostic to the particular types of relationships that are conducive to the spread of particular behaviors and ideas. We hope that our work encourages scholars in this domain to think carefully about the precise process of diffusion that may be driving network effects, which may entail accompanying, qualitative evidence gathering to better understand how real people may behave when it comes to spreading a thing in question. Not all ties are alike. The more we understand about which ties do what and when, the better our studies of diffusion and peer effects will be.

Acknowledgements
The authors are grateful for feedback from the participants of the International Conference on Complex Networks and their Applications, and the Political Networks Conference.

Author contributions
Both authors conceived and designed the study, and both wrote the manuscript. PLR led the data analysis. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The dataset analysed during the current study is from Banerjee et al. (2014) and is available in the Harvard Dataverse, doi: 1902.1/21538, at https://doi.org/10.7910/DVN/U3BIHX. The dataset used to visualize the networks in Fig. 1 is from Larson and Lewis (2016) and is available in the Harvard Dataverse, doi: 10.7910, at https://doi.org/10.7910/DVN/W1TDGZ. R code to perform the simulations is available from the authors upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 28 February 2023 Accepted: 22 April 2023
Published online: 04 May 2023

References

Alatas V, Banerjee A, Chandrasekhar AG, Hanna R, Olken BA (2016) Network structure and the aggregation of information: theory and evidence from indonesia. Am Econ Rev 106(7):1663–1704
Aral S, Van Alstyne M (2011) The diversity-bandwidth trade-off. Am J Sociol 117(1):90–171
Atwell P, Nathan NL (2022) Channels for influence or maps of behavior? a field experiment on social networks and cooperation. Am J Political Sci 66(3):696–713
Baldassarri D (2015) Cooperative networks: Altruism, group solidarity, reciprocity, and sanctioning in ugandan producer organizations. Am J Sociol 121(2):355–395
Bandiera O, Rasul I (2006) Social networks and technology adoption in northern mozambique. Econ J 116(514):869–902
Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2013) The diffusion of microfinance. Science 341(6144):1236498
Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2014) The diffusion of microfinance. Harvard Dataverse . 1902.1/21538. http://hdl.handle.net/1902.1/21538
Bearman PS, Moody J, Stovel K (2004) Chains of affection: the structure of adolescent romantic and sexual networks. Am J Sociol 110(1):44–91
Bianconi G (2018) Multilayer networks: structure and function. Oxford university press, Oxford
Boccaletti S, Bianconi G, Criado R, Del Genio CI, Gómez-Gardenes J, Romance M, Sendina-Nadal I, Wang Z, Zanin M (2014) The structure and dynamics of multilayer networks. Phys Rep 544(1):1–122
Bramoullé Y, Galeotti A, Rogers BW (2016) The Oxford Handbook of the Economics of Networks. Oxford University Press, Oxford
Brodka P, Musial K, Jankowski J (2020) Interacting spreading processes in multilayer networks: a systematic review. IEEE Access 8:10316–10341
Brummitt CD, D'Souza RM, Leicht EA (2012) Suppressing cascades of load in interdependent networks. Proc Natl Acad Sci 109(12):680–689
Burt RS (1980) Innovation as a structural interest: rethinking the impact of network position on innovation adoption. Soc Netw 2(4):327–355
Cai J, De Janvry A, Sadoulet E (2015) Social networks and the decision to insure. Am Econ J Appl Econ 7(2):81–108
Cardillo A, Gómez-Gardenes J, Zanin M, Romance M, Papo D, Pozo FD, Boccaletti S (2013) Emergence of network features from multiplexity. Sci Rep 3(1):1–6
Coleman J, Katz E, Menzel H (1957) The diffusion of an innovation among physicians. Sociometry 20(4):253–270
Cozzo E, Banos RA, Meloni S, Moreno Y (2013) Contact-based social contagion in multiplex networks. Phys Rev E 88(5):050801
Cozzo E, Kivelä M, De Domenico M, Solé A, Arenas A, Gómez S, Porter MA, Moreno Y (2013) Clustering coefficients in multiplex networks. arXiv preprint arXiv:1307.6780
Cruz C, Labonne J, Querubin P (2017) Politician family networks and electoral outcomes: evidence from the philippines. Am Econ Rev 107(10):3006–37
De Domenico M, Nicosia V, Arenas A, Latora V (2015) Structural reducibility of multilayer networks. Nat Commun 6(1):6864
Dickison ME, Magnani M, Rossi L (2016) Multilayer social networks. Cambridge University Press, Cambridge
Ferrali R, Grossman G, Platas M, Rodden J (2018) Peer effects and externalities in technology adoption: evidence from community reporting in Uganda. SSRN . https://goo.gl/NcGSvv
Gomez S, Diaz-Guilera A, Gomez-Gardenes J, Perez-Vicente CJ, Moreno Y, Arenas A (2013) Diffusion dynamics on multiplex networks. Phys Rev Lett 110(2):028701
Gondal N (2022) Multiplexity as a lens to investigate the cultural meanings of interpersonal ties. Soc Netw 68:209–217
González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Sci Rep 1(1):1–7
Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380
Greene WH (2003) Econometric analysis. Pearson Education India
Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271
Kremer M, Miguel E (2007) The illusion of sustainability. Q J Econ 122(3):1007–1065
Larson JM, Rodríguez PL (2022) Sometimes less is more: When aggregating networks masks effects. In: Complex networks and their applications, pp 214–224. Springer, Cham
Larson JM (2017) The weakness of weak ties for novel information diffusion. Appl Netw Sci 2(1):1–15
Larson JM, Lewis JI (2017) Ethnic networks. Am J Political Sci 61(2):350–364
Larson JM, Lewis JI (2020) Measuring networks in the field. Polit Sci Res Methods 8(1):123–135
Larson JM, Lewis JI, Rodríguez P (2021) From chatter to action: how social networks inform and motivate in rural uganda. British J Political Sci. https://doi.org/10.1017/S0007123421000454
Larson J, Lewis J (2016) Replication Data for: ethnic networks. Harvard Dataverse https://doi.org/10.7910/DVN/W1TDGZ
Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1361–1370

Light R, Moody J (2020) The Oxford Handbook of Social Networks. Oxford University Press, Oxford
Loken E, Gelman A (2017) Measurement error and the replication crisis. Science 355(6325):584–585
Maoz Z (2012) Preferential attachment, homophily, and the structure of international networks, 1816–2003. Confl Manag Peace Sci 29(3):341–369
Mattie H, Onnela J-P (2021) Edge overlap in weighted and directed social networks. Netw Sci 9(2):179–193
Mesch G, Talmud I (2006) The quality of online and offline relationships: the role of multiplexity and duration of social relationships. Inf Soc 22(3):137–148
Min B, Goh K-I (2014) Layer-crossing overhead and information spreading in multiplex social networks. In: APS march meeting abstracts, vol 2014, pp 17–008
Peng J, Agarwal A, Hosanagar K, Iyengar R (2018) Network overlap and content sharing on social media platforms. J Mark Res 55(4):571–585
Salehi M, Sharma R, Marzolla M, Magnani M, Siyari P, Montesi D (2015) Spreading processes in multilayer networks. IEEE Trans Netw Sci Eng 2(2):65–83
Sinclair B, McConnell M, Michelson MR (2013) Local canvassing: the efficacy of grassroots voter mobilization. Polit Commun 30(1):42–57
Solé-Ribalta A, Gómez S, Arenas A (2016) Congestion induced by the structure of multiplex networks. Phys Rev Lett 116(10):108701
Stefanski LA, Carroll RJ (1985) Covariate measurement error in logistic regression. The Annals of Statistics, 1335–1351
Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks in an online world. Proc Natl Acad Sci 107(31):13636–13641
Valente TW (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18(1):69–89
Victor JN, Montgomery AH, Lubell M (2017) The Oxford Handbook of Political Networks. Oxford University Press, Oxford
Yağan O, Gligor V (2012) Analysis of complex contagions in random multiplex networks. Phys Rev E 86(3):036103

Publisher's Note