International Journal of Research in Marketing 33 (2016) 578–599

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Sampling designs for recovering local and global characteristics of social networks☆

Peter Ebbes [a,*], Zan Huang [b], Arvind Rangaswamy [b]

[a] HEC Paris, France
[b] Penn State University, United States

ARTICLE INFO

Article history:
First received on December 6, 2013 and was under review for 6 months
Available online 19 November 2015

Guest Area Editor: Renana Peres

Keywords:
Social networks
Sampling
Subgraph sampling
Social network structure

ABSTRACT

The trajectories of social processes (e.g., peer pressure, imitation, and assimilation) that take place on social networks depend on the structure of those networks. Thus, to understand a social process or to predict the associated outcomes accurately, marketers would need good knowledge of the social network structure. However, many social networks of relevance to marketers are large, complex, or hidden, making it prohibitively expensive to map out an entire social network. Instead, marketers often need to work with a sample (i.e., a subgraph) of a social network. In this paper we evaluate the efficacy of nine different sampling methods for generating subgraphs that recover four structural characteristics of importance to marketers, namely, the distributions of degree, clustering coefficient, betweenness centrality, and closeness centrality, which are important for understanding how social network structure influences outcomes of processes that take place on the network.

Via extensive simulations, we find that sampling methods differ substantially in their ability to recover network characteristics. Traditional sampling procedures, such as random node sampling, result in poor subgraphs. When the focus is on understanding local network effects (e.g., peer influence) then forest fire sampling with a medium burn rate performs the best, i.e., it is most effective for recovering the distributions of degree and clustering coefficient. When the focus is on global network effects (e.g., speed of diffusion, identifying influential nodes, or the "multiplier" effects of network seeding), then random-walk sampling (i.e., forest-fire sampling with a low burn rate) performs the best, and it is most effective for recovering the distributions of betweenness and closeness centrality. Further, we show that accurate recovery of social network structure in a sample is important for inferring the properties of a network process, when one observes only the process in the sampled network. We validate our findings on four different real-world networks, including a Facebook network and a co-authorship network, and conclude with recommendations for practice.

1. Introduction

In recent years, there is growing interest among marketing researchers in conducting empirical research related to social networks to investigate how various network characteristics influence the processes that take place on a social network (e.g. Dover, Goldenberg, and Shapira, 2012; Goldenberg, Han, Lehmann, and Hong, 2009; Hinz, Skiera, Barrot, and Becker, 2011; Libai, Muller,

☆ The authors are listed alphabetically and contributed equally to this research. The authors gratefully acknowledge the support of the eBusiness Research Center at Penn State University. Peter Ebbes also acknowledges research support from Investissements d'Avenir (ANR-11-IDEX-0003/LabexEcodec/ANR-11-LABX-0047).

* Corresponding author at: Department of Marketing, HEC Paris, 1, rue de la Libération, 78351 Jouy-en-Josas, France
E-mail addresses: ebbes@hec.fr (P. Ebbes), zan1huang@gmail.com (Z. Huang), arvindr@psu.edu (A. Rangaswamy).

http://dx.doi.org/10.1016/j.ijresmar.2015.09.009
0167-8116/

and Peres, 2013; Liben-Nowell and Kleinberg, 2003; Nair, Manchanda, and Bhatia, 2010; Trusov, Bodapati, and Bucklin, 2010; Trusov and Rand, 2013; Watts and Dodds, 2007). These and other studies demonstrate that knowledge of network structure is important for understanding, interpreting, and forecasting processes (e.g., diffusion) that take place over networks.

Many real-world social networks of relevance to marketers are large and inherently complex, and sometimes hidden, which makes it difficult and impractical to use an entire network for empirical research. Thus, researchers typically have to work with part of a population[1] network (i.e. a subgraph) unless the study itself is conducted on a very small population network. However, relying on a subgraph for analyses necessarily raises the question of how representative the subgraph is of the population. Even if the nodes of a subgraph are representative of the nodes of the population, the links among the nodes in the subgraph may not be representative of the links among nodes in the population network. In that case, a subgraph's structure will not be representative of the population network structure, which could lead to incorrect generalizations of process outcomes observed in the subgraph. The primary objective of this research is to identify sampling approaches that result in subgraphs that are representative of the corresponding population social network. A secondary objective is to assess whether more accurate subgraphs help improve inferences about population processes that are obtained by studying those processes just on the subgraphs, or whether one could obtain accurate process metrics even without recovering accurate subgraphs.

There is a large body of literature in Statistics and Mathematics that has explored network sampling for estimating basic network properties and network totals (e.g., Wasserman and Faust, 1995, p. 34). Important work in this area was originated by Frank (see e.g. Frank, 2011, 2012 for recent reviews). Much of his work is not directly concerned with recovering the population network structure from the sampled network. Instead, the focus is on developing sampling designs that make it possible to know or estimate node inclusion probabilities, i.e., the probability that a node will be included in the sample. Knowledge of the inclusion probabilities enables a researcher to weight sample data in such a way that they accurately represent population data (e.g. Frank, 2011; Thompson, 2012, p. 174). For instance, let y_i be the value of a variable of interest for the i-th node in the population which is observed without error (e.g. y_i could be equal to 1 if node i adopted a new product and 0 if (s)he did not adopt). If we let N be the total number of units in the population, then $\tau = \sum_{i=1}^{N} y_i$ is the total number of adopters in the population. The equivalent sum in the sample ($\tau_s = \sum_{i=1}^{n_s} y_i^s$) taken only over the n_s nodes in the sample is not an unbiased estimator of τ, i.e. $E(\tau_s) \neq \tau$, where the expectation is with respect to the random sampling mechanism. This can be shown as follows. Let I_i be the sample indicator, such that $I_i = 1$ if unit i is in sample s , and 0 otherwise, and $\pi_i = P(i \in s) = E(I_i)$ is the inclusion probability of unit $i = 1, \ldots, N$. Now, $E(\tau_s) = E(\sum_{i=1}^{n_s} y_i^s) = E(\sum_{i=1}^{N} I_i y_i) = \sum_{i=1}^{N} E(I_i) y_i = \sum_{i=1}^{N} \pi_i y_i \neq \sum_{i=1}^{N} y_i = \tau$ (this bias can also be seen as a problem in scaling from the sample to the population — see Thompson, 2012; Kolaczyk, 2009). An unbiased estimator of τ may be obtained by appropriately weighting the sampled units by the inclusion probabilities, $\tau_s^w = \sum_{I=1}^{n_s} \frac{y_i^s}{\pi_i}$ (e.g. the Horvitz–Thompson estimator is a general estimator for a population total that can be used with any probability sampling plan). Calculating the inclusion probabilities π_i for general network sampling is cumbersome except for simple sampling designs such as simple random sampling where all nodes have the same known inclusion probability (e.g. Kolaczyk, 2009). Recent work based on Markov chain theory offers approximate methods for estimating inclusion probabilities for a few general designs such as snowball sampling and random-walk sampling (e.g. Frank, 2011).

Another area where knowledge of inclusion probabilities is essential is when the inference problem concerns the estimation of a probability model for a population network (see, for example, Frank, 1978; Kolaczyk, 2009; Handcock and Gile, 2010). The population network is often specified as an exponential random graph (ERGM) with certain hypothesized properties. Inference here attempts to estimate structural parameters of the population network model from a sampled network by using a weighting scheme that incorporates the node-inclusion probabilities of the sampled nodes. The weighting scheme is derived from the specified ERGM model and the selected sampling design. Here, the properties of the estimator depend on the validity of the assumed model for the population.[2] Hence, network sampling theory provides a method to estimate basic properties of population network totals and the estimation of network model parameters. The ERGM model, however, is generally not suitable as a model of social networks because social networks contain both random and non-random components — for example, family members will be connected to each other in a non-random manner.

A second stream of research in network sampling focuses on representing the structure of a *fixed or given population network*. This problem is pursued particularly in the literatures in sociology and computer science (e.g. Doreian and Woodard, 1992, Granovetter, 1976, Leskovec and Faloutsos, 2006, Reingen and Kernan, 1986). This research stream is empirical, and develops methods for generating a subgraph of a specific population network such that the subgraph is a good replica of the population network and retains its key characteristics. In a sense, a subgraph may be seen as an "avatar" of a population network, i.e., it is not an exact replica, but resembles the population network sufficiently closely in terms of its underlying structure. Here, there is only one population network (the one of interest in a study) which need not be restricted to a particular population probability model and the sampling objective is to gain information about this specific network. Thus, in this context sampling is used to

[1] We use the term population network to refer to a social network of interest in a study, and a subgraph (or sample) to refer to subset of nodes and links of the population network of interest.

[2] Thompson (2012) and Thompson and Frank (2000) consider likelihood estimation for a population network model $f(y;\theta)$. In general, inference from the sample to the larger population will be affected by the sampling procedures, because the likelihood function depends on both the sampling design and the model. However, in special situations where the sampling design does not depend on any values of the variable y (inside or outside the sample) nor on model parameters, the design is "ignorable," thereby considerably simplifying likelihood-based inference. When the sampling design is ignorable, likelihood based inference can be based just on the model likelihood. Similar observations were made by Scosyrev (2014) for estimating a population mean with unknown inclusion probabilities. He argues that the ordinary sample mean is unbiased if and only if the outcomes y are uncorrelated with the inclusion probabilities.

generate the representation of the links in a given population network. This approach to sampling is particularly useful in studies on hidden populations to recover social network structure, and in studies in which the social structure is the object of interest in itself (Thompson and Frank, 2000, p. 87). Further, subgraph sampling may be needed for reducing an observed large-scale social network to a manageable level (An, 2011, p. 527), for instance to study macro-level process outcomes for further analysis. As Granovetter (1976) notes, most social networks have had practical applications only to small groups, because any method meant to deal with groups larger than a few hundred would face insuperable obstacles. Therefore, he suggests (p. 1288) that studies involving larger groups must sample from the population network, in particular, if a study is focused on macro-level processes and outcomes.

While many empirical network studies make claims about outcomes at the level of the population, it is not clear whether the measured process outcomes from a process observed in a subgraph actually correspond to, or are representative of, population process outcomes. To the best of our knowledge, there are not many studies available that provide guidelines about the representativeness of the patterns of social relations found from a sampling procedure. And, even less is known about how marco-level process outcomes depend on the social network structure, and whether process outcomes measured on a sampled network (subgraph) are representative of population network process outcomes. Leskovec and Faloutsos (2006) make a similar observation regarding sampling from *computer* networks, where they need to match a full set of graph properties so that the sampled graphs can be used for process simulations and for complex or expensive experiments. Likewise, Libai et al. (2013) make a strong case for agent-based simulations to understand processes on a social network in the context of "seeding" programs, which require knowledge of the structure of the social network (i.e., that the network structure be observed). In both computer networks and social networks, the measured process outcomes on the sampled network should be representative of the population level outcomes.

Our study builds on the second stream of network sampling research, namely, generating a representative subgraph of a population network. In particular, in this research, we explore two important aspects of subgraph sampling that contribute to the extant literature.

First, we present a comprehensive review of a variety of sampling methods to generate subgraphs. Our study is the first to systematically and comprehensively explore how subgraphs generated from different sampling methods recover social network structure. We focus on the representativeness of the patterns of social linkages that are summarized along four main network characteristics (Granovetter, 1976): (1) node degree, (2) clustering coefficient, (3) betweenness centrality, and (4) closeness centrality. We find that traditional sampling methods (e.g. Random node sampling) perform poorly, and we need subgraph sampling methods if we are to obtain more accurate subgraphs. Importantly, we show that the best subgraph sampling method depends on the specific network characteristics that are critical for the success of the sampling effort. Specifically, different sampling approaches are required depending on whether the focus of an empirical study is on the local (i.e. node degree, clustering coefficient) or global (betweenness centrality, closeness centrality) characteristics of social networks. Given a target sample size, we find that local characteristics of the network are better recovered by subgraph sampling approaches that sample the local neighborhood around an initially selected starting node, whereas global characteristics of the network are better recovered through subgraph sampling approaches that reach nodes further away from the initially selected starting node (i.e. penetrate deeper into the network). The above results and findings can help researchers to select the most appropriate network sampling approach given the context of their study. We summarize our main recommendations in Table 2 and in Section 2.3.

Second, we demonstrate that recovering the population network structure within a sample is an important first step in correctly inferring the properties of population processes based on observations about those processes in a subgraph. Ideally, the sampled subgraph should exhibit similar process behaviors as those in the population network. We find that recovering macro flows in a subgraph depends on how well the structure of the subgraph corresponds to that of the population network. Thus, our recommendations with regard to the most effective sampling methods for a given study objective enable marketers to design and conduct experiments using subgraphs, or run (agent-based) simulations on sampled networks to gain a better understanding of how different marketing stimuli trigger various social processes in the population, and the outcomes associated with those processes.

The rest of the paper is organized as follows. In the next section we discuss related studies, current approaches in marketing for sampling social networks, and articulate our specific research objectives and measures of sampling success. We also discuss when subgraph sampling is needed and when it is not needed. In Section 3, we provide a formal description of the various subgraph sampling methods we evaluate in our study, and the simulation procedures we use to evaluate the different sampling methods. In Section 4, we present the results and findings from our simulation experiments. Finally, in Section 5, we discuss the implications of our results for theory and practice, and articulate some potential avenues for further research.

2. Background and our research objectives

2.1. Related research in other fields

To the best of our knowledge there is only one study, Leskovec and Faloutsos (2006) (L&S), that has examined the recovery of network structure in sampled networks using various popular network sampling approaches. L&S explored the characteristics of samples obtained from several known population networks for different network sampling techniques. Our study closely follows the approach taken in L&S' study especially with regard to the sampling methods studied and measures of sampling performance, but our study objectives and design choices are focused on addressing issues of relevance to social networks in marketing. For example, L&S explored how different sampling procedures perform on several known real-world networks, most of which pertain to computer-based networks (e.g., network of connections between computers on the Internet), but we focus our study on social networks which have several important characteristics that do not necessarily conform to the characteristics

of computer networks. Specifically, social networks are characterized by the following: (1) the networks are sparse (i.e., node degrees are relatively small compared to size of the network), (2) the networks have small world characteristics, i.e., distances between any two nodes in the network are small compared to the size of the network, (3) the networks have some nodes with high degrees reflecting the presence of some influential nodes (e.g., scale-free networks), and (4) the networks exhibit considerable clustering (transitivity), namely, friends of my friends are likely to also be my friends. We also note that unlike the context of computer networks used in the L&S study, in many social network applications the population network is not observed. Thus, large samples could be used in studying the L&S networks without incurring additional costs, which is not the case with social networks studied in marketing.

Unlike the L&S study which explored only known population networks, we create a range of population social networks that have parameters that are systematically varied via Monte Carlo simulations. In some sense, our study design is similar in spirit to the work of Chen, Chen, and Xiao (2013) who designed a Monte Carlo simulation to assess the efficacy of various sampling methods in recovering a network characteristic called social correlation. As explained in the next section, our simulation design allows us to carefully evaluate the factors that influence the accurate recovery of various network characteristics of interest to marketers and enables us to provide more generalizable conclusions than those provided by L&S. Thus, our recommendations are more nuanced (see Section 2.3). Specifically, even though we also find that forest fire sampling (explained in the next session) is a top performer, the appropriate burn rates overall should be much smaller as per our study, in particular when the focus is on recovering global network characteristics, such as betweenness and closeness centrality.

Although recovery of network structure using sampling approaches is an important objective on its own, we go a step further than L&S and also investigate to what extent recovery of sampled structure is necessary to obtain correct inferences regarding processes that play out in the sampled network. That is, suppose a marketer wants to track a new product diffusion process through a network (e.g. as a function of an exogenous intervention), but tracking in the population network is not feasible (e.g. the population network is unobserved or too large). Now we would need a small replica of the population network that has similar structure, so that we can track a process in the sampled network and feel assured that it would play out in a comparable way in the population network. Our results show that more accurate recovery of social structure in the sampled network does result in a process on the sampled network that more closely resembles the process on the population network.

2.2. Network sampling for marketing

Table 1 summarizes several past studies about social networks in the marketing field, their objectives, the types of networks they study, and the sampling strategy (if any) used in those studies. Collectively, these studies highlight three important factors that influence the choice of sampling strategy for social networks.

Table 1
Summary of past empirical studies in marketing that involve social networks and sampling.

Authors	Topic studied	Network size (nodes)	Type of study	Network characteristics studied	Network process studied	Network structure observed?	Network sampling strategy
Reingen and Kernan (1986)	Effect of tie strengths and subgroups	128	Empirical	Degree, path length, clusters, bridges	Referral flows	Yes	Snowball sampling
Goldenberg et al. (2009)	Role of hubs in diffusion and adoption on Cyworld	2 million +	Empirical	Degree	Adoption of items	Yes, partly	Census, random, and snowball
Trusov et al. (2010)	Log-in activity of node as a function of friends' activities	330 with 29,478 ties	Empirical	Degree	–	No	Ego sampling
Nair et al. (2010)	Prescription behavior of physicians as a function of prescription behavior of opinion leaders	1500	Empirical	Degree	Influence of immediate ties	No	Random sample of nodes
Narayan, Rao, and Saunders (2011)	Choices in a conjoint task with pre- and post-influences from peers	70	Empirical	Degree	Influence of immediate ties	No	All first-year MBA students at a school
Iyengar et al. (2011)	Adoption of a new pharma product as a function of factors, such as in-degree, self-reported leadership	193 nodes and 614 ties in total from 3 cities (SF, LA, NYC)	Empirical	Degree	Product adoption (Y/N)	Yes, partially	Cluster sampling; select entire population in clusters
Hinz et al. (2011)	Track spread of influence from a seed	1380	Experiment	Degree, betweenness	Spread of influence	Yes	Entire population of MBA students at a school
Toubia and Stephen (2013)	Content contribution on Twitter	2493	Experiment	Degree	Activity as a function of followers	No	Random sample of nodes
Libai et al. (2013)	Measuring financial performance of seeding programs	Several; max is 10,680	Empirical	Degree, clustering	Adoption dynamics	Yes	Census

First, if the focus of a study is on understanding just the effects of nodal characteristics (e.g., number of connections of a node) on outcomes associated with that node, or outcomes associated with those directly connected to that node (e.g., whether or when a node adopts a new product), then traditional simple random node sampling may often be adequate. Studies reported by Trusov et al. (2010), Iyengar, van den Bulte, and Valente (2011), Toubia and Stephen (2013), and Nair et al. (2010) fall in this category. On the other hand, if the focus of the study is on understanding the flows of influence or actions that take place via node-level activities (i.e., how micro flows lead to macro flows or cascades), then we will need sampling strategies that take into account the interdependencies between nodes and links in the network. Studies by Hinz et al. (2011), Goldenberg et al. (2009), and Libai et al. (2013) fall in this category. In such studies, there is explicit recognition of a social multiplier effect, namely, if node *A* influences node *B* and node *B* influences node *C*, then we cannot ignore the effects of node *A* on the behavior of node *C* even if A and C are not directly connected. Hence, the effectiveness of "seeding" node *A* depends on the set of links that emanate from *A* that result in a chain of linkages across the network. For example, Reingen and Kernan (1986) show that the presence or absence of subgroups in a network, as well as the nature of links between members, influence referral flows. In such studies, we require knowledge of the structure of the network (beyond the first degree connections) and traditional simple random node sampling will not be adequate for recovering that network structure. Here random node sampling is inadequate because the sampling strategy ignores the nature of interdependencies between nodes and links, and assumes independence between nodes and links, which is not consistent with the structure of social networks (our results also demonstrate this convincingly).

Second, the sampling strategy will also depend on the specific characteristics of the network that are important in a study. Whereas all studies in Table 1 provide strong empirical evidence for the importance of network structure on marketing outcomes, they focus on different structural characteristics of networks. Most studies have focused on node degree (or functions of degree) to denote network structure. Typically, these studies have used traditional sampling methods (e.g., a simple random sample of nodes, or snowball sampling) for obtaining a sample of social networks. On the other hand, Hinz et al. (2011), Dover et al. (2012), Trusov and Rand (2013), Reingen and Kernan (1986), and Goldenberg et al. (2009) provide empirical evidence that structural sociometric measures other than degree also affect marketing outcomes, such as, for example, the position of a node in a network (centrality) or the presence or absence of groups or cliques (clustering). The computation of such structural measures requires information about the network structure (i.e. the wiring diagram of how nodes are connected). Hence, following a similar reason as above, simple random node sampling would be inadequate because it ignores the nature of interdependencies between nodes and links.

Lakhina, Byers, Crovella, and Xie (2003) and Leskovec and Faloutsos (2006) point out that a single network characteristic may not be a sufficiently robust metric for characterizing the structural resemblance between the subgraph sample and the population network. They demonstrate that sampling procedures can result in subgraphs G^* that have little or no resemblance to the population network G when we consider specific network features other than aggregate network characteristics or network totals. For example, a sample degree distribution may be quite different from the population degree distribution even if aggregated counterparts (e.g., average degree) are identical. Instead, the structure of a social network may be described by multiple characteristics, and the specific set of characteristics that are of interest in one study may differ from characteristics that are of interest in another study. However, four broadly used characteristics for network structure are likely to be relevant for many different types of studies in marketing, namely, (1) node degree, (2) clustering coefficient, (3) betweenness centrality, and (4) closeness centrality (Granovetter, 1976; Van den Bulte and Wuyts, 2007). These characteristics influence processes that take place over the network, such as interpersonal influence, trust, control, or brokerage. Characteristics (1) and (2) are most associated with local network effects (e.g. clique formation), and characteristics (3) and (4) are most associated with broader network effects (e.g. network position of an individual, speed of new product diffusion).

Third, sampling strategy may depend on whether the study involves an experimental manipulation or intervention that occurs at one or more nodes in order to assess network process flows, or their effects across the entire network. There is an increasing interest in conducting experiments on social networks because of several identification challenges that arise in distinguishing between correlation and causation in survey-based studies of networks (e.g. Aral, 2015; Hinz et al., 2011; Nair et al., 2010; Trusov et al., 2010). These authors also suggest that future studies relating to social networks would benefit from field experiments.

2.3. When to do network sampling in marketing?

The previous studies in marketing highlight the importance and relevance for marketers of obtaining data on network structure. As we noted, obtaining such data is particularly beneficial when the social structure is the object of the study in itself, or when the study's objective is to track process flows through the network.

At the same time, there are marketing contexts where using a simple random sample of just the nodes may be sufficient for assuring accuracy of network inference. For example, when a study is focused on node-level properties (e.g. a node's decision to adopt or not adopt a product, the number of friends of a node, or ego-centered activities), then we do not need knowledge of network structure beyond the first degree links of the focal nodes. In such cases a random sample of just the nodes will be adequate (see, for example, Iyengar et al., 2011; Nair et al., 2010, or Trusov et al., 2010), as long as a complete list of nodes in the population (i.e. the sampling frame) is available, and detailed node-level sociometric (if relevant) and behavioral data (e.g. adoption) are observable for the sampled nodes. Likewise, if the network is of manageable size, then we may not need network sampling because we can obtain the population network by crawling. What is manageable in terms of network size depends on the context of the study, the available resources, and the computing power (e.g. An, 2011).

On the other hand, network sampling may be needed in the following cases. If the population network is large and is not observed, then network sampling is required to both understand the structure of the network, and to understand the processes

Table 2
Recommendations for network sampling (when it is needed).

Purpose of the study	Network sampling method
Understand network processes and outcomes that occur locally in a social network	Medium to high degree forest fire sampling (FF50–FF80)
Understand network processes and outcomes that occur globally in a social network	Random walk sampling
Overall outcomes and processes in social networks (i.e., without specific interest in local or global aspects)[a]	Random walk sampling

[a] This recommendation follows from our study when we perform a main effects analysis by sampling method on the average *KSD* statistic, computed across the four *KSD* statistics for degree, clustering coefficient, closeness, and betweenness. Here, we have weighted each network characteristic equally in computing the average.

that take place over that network. If the network is large and fully observed (e.g. a well-defined citation network), network sampling may still be needed depending on the specific aspects of the network or network process to be studied. Sampling is particularly relevant if the marketer is interested in understanding how certain marketer-initiated processes will play out on the network. Also, network sampling is needed when the study requires the network embeddedness of nodes to be computed from sampled data (e.g. how central are the nodes in the network).

Once we determine that network sampling is required in a particular study, then our sampling recommendations summarized in Table 2 are useful in selecting a sampling method for the study. We will elaborate on our findings and recommendations in the remainder of the paper.

2.4. Research objectives of this study

The primary purpose of our research is to assess the efficacy of both traditional and subgraph sampling methods in recovering the underlying structural characteristics of population networks. Define a population network as $G = (V, E)$, where $V = \{v_1, \ldots, v_N\}$ and $E = \{(v_i, v_j)\}$ are the node set and link set, respectively. Then a subgraph of G is defined as $G^* = (V^*, E^*)$, where the (sampled) node set and link set of graph G^* are $V^* \subseteq V$ and $E^* \subseteq E$, and $(v_1, v_2) \in E^* \rightarrow v_1, v_2 \in V^*$. The *traditional* sampling procedures such as random node sampling when applied to networks, obtain a subgraph of a given size (as denoted by the size of V^*) by randomly selecting nodes from V and then selecting post-hoc the links from E that are associated with all those nodes. Subgraph sampling differs from traditional sampling, because in subgraph sampling the sampling procedures take into account the relationship between nodes and their links; that is, nodes and links are sampled jointly. Both traditional and subgraph sampling approaches result in an *induced subgraph* G^* of the population network. In marketing studies, E does not have to be observed for these sampling approaches to be useful, because survey methods may be used to discover the links between the nodes (e.g. Reingen and Kernan, 1986).

We now provide brief descriptions of the nine different sampling methods we include in our study, which are summarized in Table 3.

2.4.1. Random-node sampling

We include two variations under this category: (1) *uniform random node method* (RN) under which we select the subset of nodes V^* independently and with equal probability from V, and the link set E^* is obtained post-hoc from the links that connect the selected nodes (See, Frank, 1978), and (2) *degree-based random node method* (DRN) (Leskovec and Faloutsos, 2006) in which we select nodes independently and with a probability proportional to their degree. Specifically, we select one node at a time from the unselected nodes with node v_i being selected with a probability equal to $d(v_i)/\sum_{v \text{ in unselected nodes}} d(v)$ until the desired number of nodes has been selected ($d(.)$ denotes degree). We note that the DRN method is of little practical value when the population network is unknown, particularly, its degree distribution. Nevertheless, this method provides a useful benchmark for our research. We also note that RN for subgraph sampling is different from randomly sampling just the nodes (e.g. Section 2.3).[3] In RN for subgraph generation, we take a random sample of nodes (V^*) and then include all links (E^*) between the selected nodes to build the subgraph $G^* = (V^*, E^*)$. When selecting a random sample of just the nodes, we do not include the link set E^* in the study and only work with information associated with the nodes in V^*. The subgraph G^* would be needed, for instance, if a marketer is interested in conducting process simulations or other interventions within the sampled graph (see Section 2.3).

2.4.2. Random-edge method (RE)

In this method (also known as incident subgraph sampling), we independently and with equal probability select one edge (v_i, v_j) at a time from E and include v_i and v_j into the set of selected nodes, until the number of nodes correspond to the required sample size. This method is often not practically useful because we would need to know all edges in the population network to implement the procedure.

[3] Kolaczyk (2009; pp. 124–145) discusses the differences between these two random-node sampling approaches when the goal is to estimate the average degree in the population. Whereas the average sample degree from randomly sampling just the nodes is unbiased for the average population degree, the average degree in the induced subgraph sampled using the RN method is biased. This bias can potentially be remedied by using a Horvitz–Thompson estimator, which in this particular case, however, would require knowledge of the exact number of nodes in the population. Also, randomly sampling just the nodes is similar to "unlabeled star sampling" in Kolaczyk (2009; sec. 5.3), whereas our definition of RN is similar to his concept of "induced subgraph sampling."

Table 3
Subgraph sampling approaches considered in this study.

Sampling method		Brief description[a]
Uniform random node	RN	Select nodes with equal probability; post-hoc includes the links that connect the selected nodes.
Degree based random node	DRN	Select nodes with probability proportional to their degree; post-hoc includes the links that connect the selected nodes
Random edge	RE	Select edges with equal probability; post-hoc includes the nodes belonging to each selected edge.[b]
Ego centric	EGO	Select nodes with equal probability and include all their alters in the sample.[b]
Snowball	SB	Select one (or more) starting nodes at random, select the entire set of alters in the sample, and repeat this process for each included alter.[b]
Random walk	RW	Select one (or more) starting nodes at random, select randomly one alter from the entire set of alters in the sample, and repeat this process for the selected alter.[b]
Forest fire	FF	Select one (or more) starting nodes at random, select randomly *X*% alters of the entire set of alters in the sample, and repeat this process for each included alter.[b] In this study we set *X* to 20, 50, and 80.

[a] For each sampling method, the sampling process is stopped until a preset sample size of number of nodes is reached.
[b] In a second step, we also include all edges in the sample that connect any pair of included nodes.

2.4.3. Egocentric method (EGO)

The *egocentric method*, also known as labeled star sampling, is motivated by the egocentric network data commonly used in sociology (Capobianco and Frank, 1982). We first select a seed node *v* (called ego) uniformly at random and then include *v* and all its neighbors (called alters) in the sample, and then continue to select a new ego (at random) and its alters until we reach the desired number of nodes. When the sample node size is exceeded by including all alters of the last ego, we randomly choose a portion of these alters such that the exact sample size is obtained.

2.4.4. Subgraph sampling methods

Under the methods in this category, we include nodes in the sample by navigating (i.e., exploring) the population network following the edges from a randomly selected starting seed node *s*. From *s*, we select either a subset or the entire set of the unselected neighbors of *s*, and then repeat this process for the selected neighbors. If the desired number of nodes cannot be reached (i.e., the exploration process falls into a component of the population graph isolated from others) we select a new random seed node to restart the exploration process.

There are several variations of such "graph exploration" methods depending on how we select the neighbors of a node *v* that are reached via the exploration process. The *snowball method* (SB) includes all unselected neighbors of *v*. This method is quite popular in sociology (Frank, 1979; Goodman, 1961). The *random walk method* (RW) (e.g., Klovdahl, 1977) selects exactly one neighbor uniformly at random from all the unselected neighbors. The snowball and random walk methods represent two extremes of a continuum (i.e., a continuum ranging from selecting just one, or all, of the immediate neighboring nodes), and are also known as breadth-first and depth-first search respectively. Between these two extremes lies the *forest fire method* (FF) (Leskovec and Faloutsos, 2006), which selects *l* unselected neighbors of *v* uniformly randomly where the number of neighbors selected at each stage *l* is set to be a given percentage (i.e., the *burn probability*) of the remaining unselected neighbors.

To obtain insights regarding the effect of burn probability on the quality of the resulting sample network we set the "burn probability" percentage to 80% (FF80), 50% (FF50), and 20% (FF20). To ensure that our forest fire method falls between the random walk and snowball methods, we take the minimum number of neighbors to be selected at each stage equal to 1 (if at least one edge is available). Therefore, the effective burn probabilities of our sampling procedures are slightly larger than the specified percentages, especially when small-degree nodes are common in the network. For the RW and FF methods we allow for the possibility that the sampling process will revisit nodes that have already been included into the sample (although those nodes do not appear multiple times in the sample).

For all sampling procedures discussed above, we apply a two-step process to determine the nodes and edges included in the sample subgraph: (1) identify the set of nodes to be included in the sample through either random selection of nodes/edges, or based on graph exploration and (2) include *all* edges among these nodes. Whereas it is common in the literature to only do step (1) for including the edges in the sample network, including all edges among the sampled nodes in a second step will improve sampling performance. The second step is intuitive in the context of selecting sample *social* networks, because we can obtain from a survey of each respondent in the sample his or her links to everyone else already in the sample, unless the sample becomes very large. In all our sampling procedures, we sample without replacement, that is, each node appears only once in the sample.

3. Simulation procedure to assess different sampling methods

To assess the quality of the different sampling methods, we compare the distributions of node degree, clustering coefficient, betweenness centrality, and closeness centrality in the sampled networks with the corresponding distributions in the population network. We first briefly describe the four network characteristics and their relevance for marketing (see also Van den Bulte and Wuyts, 2007), and present the mathematical details of their computation in Appendix I. We then discuss the networks used and the measures of sample quality.

3.1. Network characteristics studied

Degree and *clustering* are most associated with local network effects (e.g. peer influence). The *degree* of a node in a network is the number of edges connected to that node, which is a measure of the local influence which a node can exert, or the influence that other nodes connected to a focal node can exert on that node. *Clustering* is a measure of a network's transitivity, i.e., a measure of the extent to which the connections of a node are also connected. The degree and clustering coefficient associated with a node are unlikely to be good measures of network-level influence of that node. For example, a node connected to other influential nodes (e.g., nodes that have higher degree) can have a greater influence on network process outcomes compared to another node of equal degree connected to less influential nodes. Interestingly, local influence is likely to be greater in the latter case, because the other nodes are more dependent on the focal node. To capture these types of characteristics, we consider two measures of node "centrality". Centrality reflects the node's position in a network, and goes beyond the immediate local network. We study two versions of node-based centrality measures derived from the path length information (Albert and Barabási, 2002). First, *betweenness centrality* captures the idea that a well-placed node that lies on the shortest paths between other nodes has the potential to control communication in the network and command attention (Freeman, 1979). Second, *closenesses centrality* can be viewed as an index of the efficiency with which a process starting at that node can reach other nodes, and is a summary indicator of a node's access to information and other resources in the network. In a network with larger values of closeness centrality, diffusion of information can be fast (other things equal).

3.2. Networks used for analysis

For our study, we consider two distinct population network models, namely, a small-world and a modified power-law model. These networks possess properties that are associated with social networks, namely, relatively short path length and high clustering (Albert and Barabási, 2002, Levinea and Kurzbanb, 2006, Strogatz, 2001), and are thus good approximations to a large number of real social networks. The two network models differ in terms of their degree distributions, having either a Poisson degree or a power-law degree distribution.

Watts–Strogatz networks have short path lengths, a large clustering coefficient (compared to Bernoulli random graphs), and an approximately Poisson degree distribution. We denote these networks as WS in the rest of the paper. A key property of WS networks is that most nodes have a degree not too far from the average degree. We obtain the population network for the WS network by starting with a regular network in which each node is connected to a fixed number of neighboring nodes and applying a random rewiring process to that network (Watts and Strogatz, 1998). In power-law networks, on the other hand, some nodes have very large degrees and most nodes have just a few connections (i.e., a small degree). Power-law degree distributions have been observed in a wide range of networks (e.g. Albert and Barabási, 2002, Newman, 2001, Strogatz, 2001), but unlike WS networks, power-law networks may or may not have a high clustering coefficient. However, most social networks that exhibit power law behavior also have relatively high clustering (Mislove, Marcon, Gummadi, Druschel, and Bhattacharjee, 2007). To generate social networks that have a power-law degree distribution and also exhibit clustering, we use the approach proposed by Holme and Kim (2002), which produces networks for which the degree distribution is power-law, but which allows tunable levels of clustering for the network — we denote such networks as HK in this paper. Comparing WS and HK networks allows us to investigate whether the presence or absence of high degree nodes have an effect on the ability of various sampling methods to recover key properties of the population networks. For instance, the local structure of a network near a high degree node is likely to be different from the structure farther away from that node, and when high degree nodes are oversampled (e.g. snowball sampling is known for oversampling high degree nodes because of their high connectivity), the sample network would have low resemblance to the population network.

In generating both WS and HK networks, we have chosen parameter levels such that the simulated networks have levels of clustering (ranging from 0.09 to 0.25) that resemble those of large-scale social networks found in practice (e.g. Levinea and Kurzbanb, 2006, Mislove et al., 2007, Valente, Watkins, Jato, Van der Straten, and Tsitsol, 1997). In particular, we manipulate the density of the networks (defined by the number of edges) at four levels (low, medium–low, medium–high, and high). Dense networks are inherently more complex in the regions local to a node. Therefore, other things equal, we should expect that the quality of sampled subgraphs to decrease as the population network density increases, particularly for recovering clustering coefficient. Appendix II summarizes the technical details about how we generated the different population networks to have the desired characteristics.

In addition, we consider two sizes for the population networks, namely 10,000 and 20,000 nodes. This allows us to investigate whether the findings are sensitive to sample size in reference to the size of the population. This issue may be particularly relevant when degree distributions are long-tailed as in HK networks, as typically a lot of data is required to capture the tails adequately. Thus, for our main analysis we have 16 simulated population networks: 2 (population structures, namely, WS and HK) $\times$ 4 (node density levels) $\times$ 2 (population sizes). Finally, to assess the robustness and validity of our results we use two more simulated networks and two empirical networks. Specifically, we consider two simulated networks that are theoretically closely related to the WS and HK networks. The first is the rewired connected caveman (CC) network (Watts, 1999), which starts with a connected caveman model with fully-connected sub-networks of size m ("caves") arranged in a ring structure by using one edge from each cave to connect to another cave. The CC network is similar to the WS network characterized by high clustering, short path length, and Poisson degree distribution. The CC network differs from the WS network in that the caves represent a tighter local community structure. The second is a network based on Preferential Attachment (PA) model (Barabási and Albert, 1999). PA

networks are similar to HK networks except in terms of having lower clustering coefficients. For validating our results with CC and PA, we also varied the density levels, so that we have 8 population networks for validation: 2 (population structures, namely, CC and PA) × 4 (node density levels).

We obtained two real-world social networks for validation, namely, a Facebook friendship network of students at a large public university graduating in 2010 and an academic network containing co-authorship relationships. We adapted the Facebook API to construct the friendship network. Our data only contains public Facebook profiles at the time of the data collection. For constructing our co-authorship network, we used the high-energy particle physics co-authorship database,[4] where we include a link between two authors if they have co-authored at least three papers. We used this criterion to ensure that this network has a degree density that falls within the range used in our simulations.

Table 4 summarizes the characteristics of the population networks in terms of basic statistics about network size and density, and also the average values of node degree, clustering coefficient, betweenness centrality, and closeness centrality. Where applicable, we also report the estimated power law distribution exponent.

3.3. Measures of sample quality

We compare *complete* distributions for these four network characteristics in the sample versus the population to assess the efficacy of different sampling methods (instead of, for example, comparing only aggregate network metrics, such as average degree). There are several reasons for this choice. (1) There are no analytic formulations for the distributions of the characteristics of complex networks of the type explored in this study, i.e., it is not possible to analytically derive the sampling distribution for a subgraph statistic (e.g., average degree) as a function of a sampling method, as reported, for example, in Salganik and Heckathorn (2004), Thompson (2006), and Kolaczyk (2009). To generate such analytic formulations, we would at the minimum need to assign appropriate probability weights to the values observed at each node in the sample, but we do not have general tractable probability models to encapsulate the variety of social networks that could be observed in practice. (2) The distributional shapes of network characteristics contain information about network structure. Dover et al. (2012) find that the speed of contagion depends on the shape of the degree *distribution* of the underlying network, while Watts and Dodds (2007) and Reingen and Kernan (1986) provide evidence that the *relative* presence of hubs and bridges affect how influence proceeds from micro flow to a macro flow. Also, an aggregated network characteristic may not adequately characterize the structural resemblance between a sample and population network (Lakhina et al., 2003; Leskovec and Faloutsos, 2006). (3) By comparing complete distributions we can potentially detect sampling biases that occur because of the incomplete data in subgraphs, i.e., the comparison incorporates non-ignorable missing data (e.g., Little and Rubin, 2002).

To assess the goodness-of-fit between distributions of a population network characteristic $j = 1,2,3,4$ and that of the corresponding sample values, we use the Kolmogorov–Smirnov D-statistic (*KSD*), defined as $D_j = \max_{\text{all } x} |F_j(x) - F'_j(x)|$ where x is over the range of the characteristic; F_j and F_j' are the empirical cumulative distribution functions of the population network and sample network, respectively.[5] Note that we do not use the *KSD* to test whether the distribution of the sample network characteristics is from the same distribution as the population network. We only use it as a measure of goodness-of-fit (or discrepancy) between the sample and population distributions. It is clear that smaller values of D_j indicate superior fit of the sample distribution with the population distribution. In addition to the *KSD*, we considered the following "relative" measure of cross-entropy for comparing discrete distributions:

$$H(p,q) = \frac{\sum_x p(x)\log(q(x)) - \sum_x p(x)\log(p(x))}{\sum_x p(x)\log(p(x))} \tag{5}$$

where $p(x)$ is the distribution of a characteristic in the population and $q(x)$ is the distribution of that characteristic in the sample. $H(p,q) = 0$ when the two distributions are identical. We note that *KSD* is specified as a function of the cumulative distribution and the cross-entropy is a function of the density distribution.

3.4. Mini-populations: benchmarks for performance

We propose the concept of a "mini population" to help us establish a benchmark for assessing the performance of our sampling methods. Mini-populations also enable us to investigate the potential effects of scaling of population characteristics induced by sampling, and the efficacy of our standardization procedure to reduce scaling effects. There could be an intrinsic "scaling problem" in computing the *KSD* and cross-entropy if the distribution of a network characteristic is a (non-linear) function of the number of nodes. For example, the degree of a node is likely to be larger in the population than in the sample because the population network has more nodes. Likewise, characteristics such as betweenness and closeness centrality are affected by path lengths, which increase with network size. However, clustering coefficient is mainly a local measure which is likely not substantially affected by changes in network size.

[4] Available at https://kdl.cs.umass.edu/display/public/HEP-Th (last accessed: August 2015).
[5] See Lindgren (1993) for a discussion on the usefulness of the D-statistic for both continuous and discrete distributions.

Table 4
Characteristics of population networks included in our study.

Graph	Graph generation parameters	Density	# nodes	# edges	Avg. degree	Power law exponent	Avg. clustering coefficient	Avg. betweenness centrality (×0.001)	Avg. closeness centrality
Holme–Kim	$p = 0.4$ (probability of adding a triangle after adding a random edge)	Low	10,000	19,980	3.9960	2.7693	0.1525	0.4032	0.2010
		Medium	10,000	39,944	7.9888	2.9049	0.1691	0.2960	0.2542
		Dense	10,000	49,912	9.9824	2.9084	0.1377	0.2705	0.2713
		High	10,000	79,739	15.9478	2.9430	0.0923	0.2256	0.3086
		Low	20,000	39,992	3.9992	2.8333	0.2999	0.2275	0.1820
		Medium	20,000	79,936	7.9936	2.9183	0.1660	0.1585	0.2413
		Dense	20,000	99,888	9.9888	2.9330	0.1365	0.1450	0.2577
		High	20,000	159,703	15.9703	2.9433	0.0886	0.1223	0.2914
Watts–Strogatz	$p = 0.3$ (probability of rewiring each edge)	Low	10,000	20,000	4		0.1859	0.7554	0.1172
		Medium	10,000	40,000	8		0.2230	0.4378	0.1861
		Dense	10,000	50,000	10		0.2360	0.3846	0.2065
		High	10,000	80,000	16		0.2414	0.3000	0.2502
		Low	20,000	40,000	4		0.1831	0.4110	0.1087
		Medium	20,000	80,000	8		0.2240	0.2395	0.1728
		Dense	20,000	100,000	10		0.2320	0.2104	0.1921
		High	20,000	160,000	16		0.2414	0.1661	0.2315
Barabasi–Albert		Low	10,000	19,996	3.9992	2.8313	0.0053	0.4005	0.2017
		Medium	10,000	39,984	7.9968	2.8640	0.0070	0.2869	0.2600
		Dense	10,000	49,975	9.9950	2.9382	0.0067	0.2671	0.2737
		High	10,000	79,936	15.9872	2.9664	0.0094	0.2241	0.3099
Rewired connected caveman	$p = 0.3$ (probability of rewiring each edge)	Low	9995	19,997	4.0014		0.2655	0.7090	0.1156
		Medium	9999	39,996	8		0.3320	0.7683	0.1828
		Dense	9999	49,995	10		0.3388	0.4480	0.2038
		High	9996	79,968	16		0.3483	0.3913	0.2474
Facebook			7227	56,384	15.6037	3.0372	0.1415	0.3047	0.2602
Coauthor			2134	3700	3.4677	3.2596	0.3598	0.4031	0.1286

We generated 25 smaller "mini-population" networks for each of the eight conditions for HK and WS population network listed in Table 4. The sizes of these mini-population networks were 500, 1000, and 2000, nodes, and they were constructed *using the same graph-generation mechanism used for generating the population networks*. In other words, these mini-populations are smaller replicas of the (larger) population network. We then compared the distributions of the four network characteristics (i.e., degree, clustering coefficient, closeness and betweenness) of each of the 25 mini-populations to the distribution of the corresponding population network using the *KSD* and the relative cross-entropy measure.

Table 5
Kolmogorov–Smirnov D-statistics for mini-populations and main population.

	Density (avg. degree)	# nodes	Clustering coefficient	Degree	Betweenness centrality	Closeness centrality
HK	Low (4)	500	.067	.019	.111	.052
		1000	.042	.014	.102	.046
		2000	.023	.010	.107	.040
	Medium–low (8)	500	.123	.024	.038	.056
		1000	.075	.018	.029	.042
		2000	.037	.012	.025	.032
	Medium–high (10)	500	.209	.021	.036	.072
		1000	.130	.015	.029	.052
		2000	.072	.012	.021	.033
	High (16)	500	.461	.060	.042	.122
		1000	.291	.024	.034	.124
		2000	.161	.013	.025	.094
WS	Low (4)	500	.042	.021	.036	.033
		1000	.027	.012	.024	.023
		2000	.019	.008	.022	.017
	Medium–low (8)	500	.055	.019	.030	.029
		1000	.043	.012	.025	.022
		2000	.025	.010	.017	.016
	Medium–high (10)	500	.054	.019	.030	.033
		1000	.031	.014	.025	.019
		2000	.023	.011	.019	.017
	High (16)	500	.091	.029	.030	.034
		1000	.050	.016	.021	.023
		2000	.030	.011	.016	.028
	Average		.091	.018	.037	.044

Notes: Degree, and betweenness and closeness centrality were standardized before computing the *KSD*.

A common practice in applied statistics to account for scaling is to bring the data from the different sets to a common scale through standardization by subtracting a measure of location and dividing by a measure of scale, allowing for a comparison of the underlying characteristics. For our study, an important characteristic is the *shape* of the distribution (Lakhina et al., 2003), so that, for example, we can determine whether the sample node degree has a power-law tail distribution in the case of the HK model or a symmetric distribution in the case of a WS model. Table 5 presents the average values of D_j (across the 25 mini-populations) for each graph characteristic $j = 1, 2, 3, 4$ (standard deviations are generally smaller than 0.030). Overall, the average *KSD* across the four characteristic with standardization is 0.047, and without standardization (results not shown), the average *KSD* is substantially higher at 0.55. We find similar results for the relative-cross entropy measures (results not shown), where the average relative cross-entropy measure is 0.12 with standardization and 1.01 without standardization. This result confirms that standardization is important in comparing the population and sample networks, because the mini-population networks are theoretically similar to the population network, and an appropriate measure of correspondence should reveal that fact. Consistent with this argument, we find that distributions of degree, betweenness centrality and closeness centrality of the smaller mini-populations are very similar to the corresponding distributions in the population, and scaling as a function of sample size, is largely removed *after* standardization. We find very similar insights using the *KSD* and the relative cross-entropy measures for the mini-population analysis. However, *KSD* is easier to compute because it is obtained from the cumulative distribution directly, whereas the cross-entropy measure requires binning, particularly for the continuous centrality measures. For the very skewed power-law tailed distributions we also found that standardization was sensitive to outliers and we used the difference between the third and first quartile for the scale and the average of the first and third quartile for the location measure in our standardization procedure. Based on our analysis with mini-populations, we standardize the measures for degree, betweenness centrality, and closeness centrality, before comparing the sample distribution with the population distribution.

The *KSD* values in Table 5 provide an overall benchmark against which the *KSD* from the samples generated from various sampling methods may be compared. These "optimal" *KSDs* based on the mini-populations ranged, on average, from 0.018 to 0.091. We expect that subgraphs generated via any sampling method will have higher *KSDs*, on average, than these benchmarks which were generated directly from the population graph-generating mechanism. Further, in the cases of small samples from networks that have high density, we see that the mini-populations are less similar to the population than in other cases, particularly for the HK networks. Thus, the density of a population network is likely to be an important factor in generating good quality samples. In our simulation experiments, we manipulate both sample sizes and density levels.

4. Results and findings

The base case for our simulation contains for each type of population networks (HK and WS) four density levels and 10,000 or 20,000 nodes in the network (see Table 4). We used the nine sampling methods to each draw 30 samples for each of the following sample sizes: 100, 200, 300, 400, 500, 700, 1000, and 2000. Hence, we obtained a total of 34,560 ($=2 \times 4 \times 2 \times 9 \times 8 \times 30$) sampled networks for the base analysis. For the four networks we use for validation (CC, PA, Facebook, and Co-authorship), we used the same sample sizes.

4.1. Main effects of simulated factors

To understand the overall statistical effects of the simulation design factors on the average *KSD*, we first performed an ANOVA using the five main factors (population network model, population size, density of network, sampling method, and sample size) and all interactions. We summarize the results in Table 6. We only highlight main effects and interaction effects for which partial η^2 values are greater than 0.2.

We can see that the degree recovery is the least affected by the simulation factors, which is consistent with our expectations, and our findings with respect to the mini populations. And, betweenness centrality is the characteristic most influenced by the

Table 6
ANOVA for effects of simulation parameters.

Source	df	Partial η^2 (clustering coefficient)	Partial η^2 (degree)	Partial η^2 (betweenness centrality)	Partial η^2 (closeness centrality)
Density	3	0.856	–	0.413	–
Sample size	7	0.505	–	0.693	0.352
Sampling method	5	0.202	–	0.746	0.582
Population size	1	–	–	0.284	–
Network model × density	3	0.458	0.369	0.398	–
Network model × sampling method	5	0.376	–	0.441	0.315
Sampling method × density	15	0.341	–	0.359	0.235
Sampling method × sample size	35	–	–	0.245	0.251
Sampling method × network model × density	15	–	–	0.245	–
η^2 of model	767	0.910	0.697	0.913	0.717

Notes: (1) Only considers subgraph sampling methods including EGO networks, and ignores traditional sampling based on RN or RE selection, which all performed substantially worse than the subgraph sampling methods. (2) Only reports effects for which $\eta^2 > 0.2$. (3) All effects were significant for $\alpha = 0.10$ except the highest-order interaction for clustering coefficient.

simulation factors. We also find that the sampling method is more important for recovery of global characteristics (e.g., centrality) than the local characteristics of a network (e.g., degree and clustering). Next, we examine the main effects of two simulation factors, namely, sample size and population degree density on recovery of the four network characteristics. As expected larger samples lead to improved recovery. What is surprising is that sample size has only a modest effect on the performance of the subgraph sampling methods, as shown in Fig. 1A. Sample size has the most substantial effects on the recovery of betweenness centrality, and has the least impact on recovery of degree distribution. Fig. 1B shows that the population density has a strong main effect on the recovery of the cluster coefficient distribution, where the recovery is worse for more dense population networks. Below, we explore the effects of sample size and density on the differential performance of the alternative sampling methods in more detail.

Interestingly, population size has only a minimal impact on recovery of the population network structure. As the size of the population increases from 10,000 to 20,000 nodes while keeping the absolute sample size the same, the corresponding *KSD*s shift upward in almost a parallel fashion for both HK and WS networks. This holds across the simulated conditions. Thus, our recommendations regarding the appropriate sampling methods are unaffected by increases in population size, although we can expect a degradation in the performance of all sampling methods as population size increases (other things equal).

The ANOVA results also show that there is no one best sampling method across the different network characteristics. Rather the best sampling method depends on the nature of the population network (sampling method × network model interaction), and network density (sampling method × density), as well as the specific network characteristic of interest. Next, we elaborate on the effects of the simulation parameters on each of the four network characteristics. We will first discuss the results for clustering coefficient, followed by degree, and betweenness and closeness. The main results are summarized in Figs. 2–5 and Table 7.

4.1.1. Recovery of clustering coefficient distribution

Fig. 2 shows the mean *KSD*s and 95% confidence intervals for the nine sampling methods for the clustering coefficient. Each mean value is computed from 3840 samples for the 4 density levels, 2 network types, 2 population sizes, 8 sample sizes, and 30 sample draws. We can readily see that the traditional sampling methods (RN, RE, DRN) have very poor performance, with mean *KSD* of 0.930, 0.728 and 0.816, respectively, with an overall mean of 0.825 across these three traditional sampling methods. On the other hand, collectively, the subgraph sampling methods have a mean *KSD* of 0.391.

Among the subgraph sampling methods, FF50 and FF80 perform the best, with the two of them having an overlapping confidence interval, followed by RW and SB. EGO and FF20 performed the worst among the subgraph sampling methods. As we only include five burn probability levels in our study (RW, FF20, FF50, FF80, and SB), our recommendations regarding burn probability are only broadly indicative of superior performance. In particular, for the clustering coefficient, our general recommendation would be a burn probability in the range of 40% to 80%. However, for purposes of validating our recommendations, we will fix the burn probability at 50% and use FF50 as our recommended method.

For assessing the validity of our recommendations, we benchmark our recommended method with the four sampling methods commonly employed in the literature, namely, RN, EGO, RW, and SB. We define an *underperformance measure* of a sampling method k as:

$$\frac{\text{D-value of method } k}{\text{D-value of recommended method}} - 1. \tag{6}$$

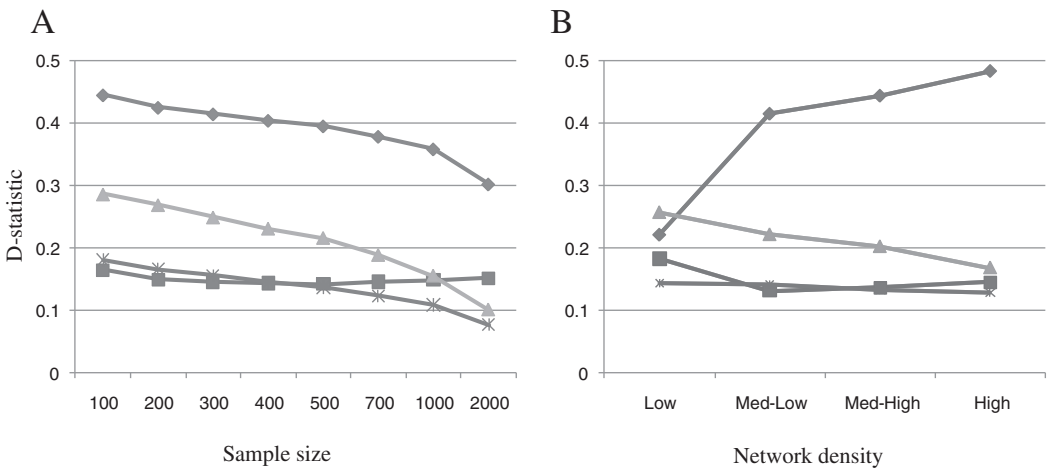

Fig. 1. A and B: Main effect of sample size (left) and density (right) on the average performance (D-statistic) of subgraph sampling methods. Here a diamond indicates cluster coefficient, a square indicates degree, a triangle indicates betweenness centrality, and a star indicates closeness centrality.

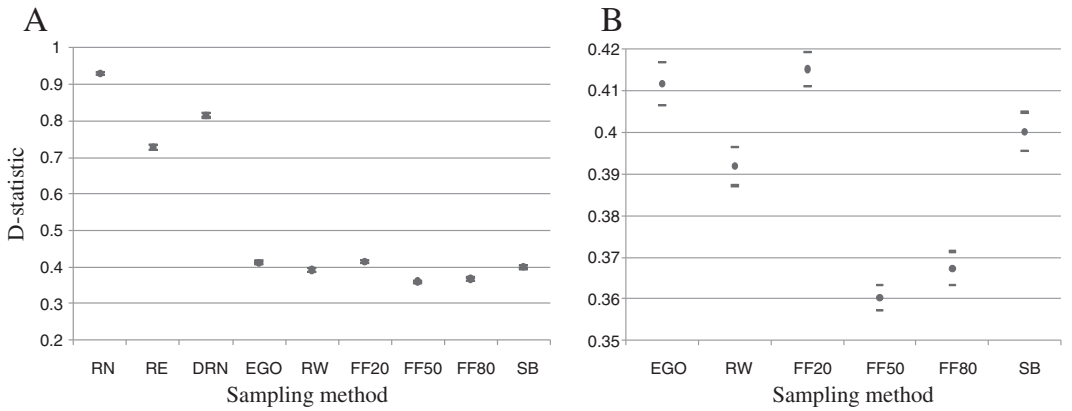

Fig. 2. Clustering coefficient D-statistic and the confidence interval by sampling method.

Table 7 shows *KSD* and underperformance measures of the recommended method (FF50) and the four benchmark methods for our four validation networks (PA, CC, Facebook, and Coauthorship). For completeness, this table also shows information for the HK/WS networks from which we developed our recommendations.

Our validation analysis shows that, overall, the recommended FF50 sampling improves recovery of clustering coefficient distribution for our four validation networks, when compared to the recovery from RN, SB, RW, and EGO sampling methods, with an average improvement of 109% (with respect to the underperformance measure). Consistent with our main analysis, the validation networks also show poor performance of traditional sampling methods, represented by RN having large underperformance measures.

4.1.2. Recovery of degree distribution

Fig. 3 summarizes the overall mean *KSD* for the various sampling methods for the degree distribution. Again, it is clear that the traditional sampling methods (RN, RE, DRN) perform poorly compared to the subgraph sampling methods. All the subgraph sampling methods perform approximately equally well, with FF80 performing best overall with an average *KSD* of 0.138. For most practical purposes, it appears that any of the subgraph sampling methods may be used for determining degree distribution.

Again, we validate our recommended method (FF80) on the four validation networks against RN, EGO, RW, and SB, and these methods had an overall underperformance measure of 42.1%, on average (Table 7). FF80's improvement is largest over RN and RW. This result with respect to RN is important — it suggests that even for recovering just nodal characteristic, such as degree distribution, subgraph sampling methods improve performance over traditional sampling. The reason for this superior performance is that the connectivity of nodes within the subgraph sample obtained with FF80 is more representative of the population network than is the connectivity among the nodes in the sample generated by RN sampling (see also footnote 4).

Both clustering coefficient and degree are structural elements of networks that primarily influence local flows. We find for both clustering and degree that FF sampling with moderate to high burn probability recovers these distributions best, possibly because such a burn value allows for a local exploration of the network, without oversampling locally, as may be the case with SB sampling. The next set of results shows that when the interest is on the network-level (global) influence of nodes, a deeper exploration of the network is important, and therefore, subgraph sampling methods that go further into the network are required.

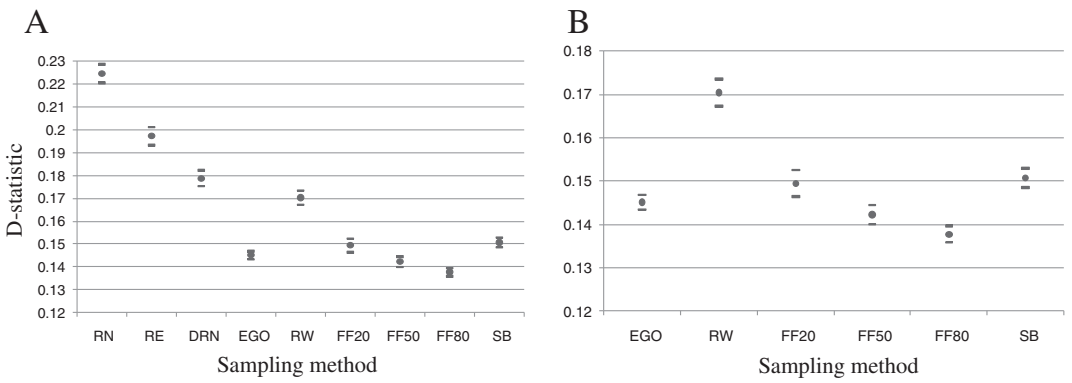

Fig. 3. Degree D-statistic and its confidence interval by sampling method.

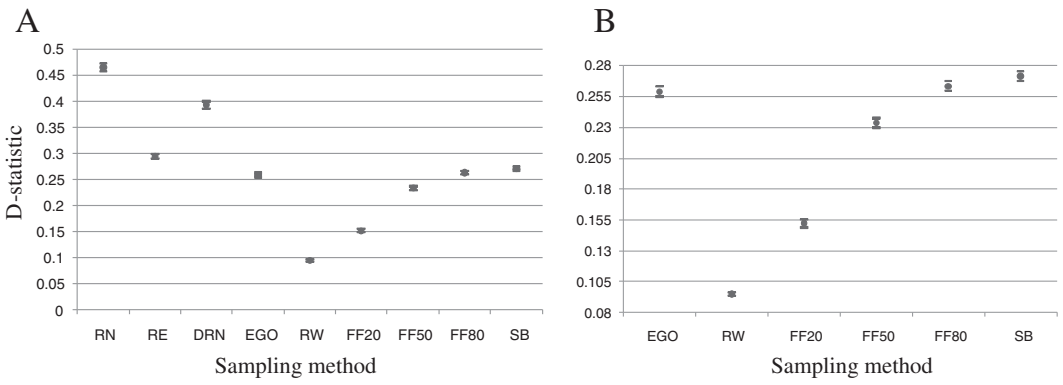

Fig. 4. Betweenness D-statistic and its confidence interval by sampling method.

4.1.3. Recovery of betweenness and closeness centrality distributions

The results for betweenness and closeness centrality are similar and we discuss them together. Figs. 4 and 5 summarize the overall means of the *KSD* for the various sampling methods. The ANOVA model discussed above indicates that sampling method is the most important factor influencing recovery of the population network characteristics. Again, the traditional sampling methods perform poorly compared to the subgraph sampling methods. We can also see that recovery becomes poorer as the burn probability increases (i.e., the best method is RW).

As a validation check of our recommendation, we compare the performance of RW to the other extant methods (RN, EGO, and SB). The combined average improvement in *KSD* for betweenness and closeness centrality across the four holdout networks is 139.8% (Table 7). The improvement is largest over RN (305%) and smallest over SB (52%), on average. We note, however, that for the Facebook network, the SB and EGO sampling approaches obtain a better recovery of betweenness than RW, with *KSD* that are 29.8% and 23.2% lower, respectively than our recommended method, RW. In fact, for the Facebook network, FF sampling with low burn (FF20) is the best (average $KSD = 0.101$). This result suggests that there may be a structural difference between the Facebook network and HK/WS networks, which affect recovery of betweenness centrality.[6]

For closeness centrality, the performance of RW sampling is robust across the four validation networks, and offers consistent improvement in *KSD*, with an average improvement of 232% over the extant methods RN, EGO, and SB. The improvement is largest over RN (492%), followed by SB (123%) and EGO (84%).

In sum, our main analysis lead us to two conclusions that were robust across the four validation networks: (1) subgraph sampling methods dominate traditional methods in recovering important characteristics of social networks; and (2) local characteristics of the network (e.g. degree and clustering) are better recovered by higher burn probabilities, whereas global characteristics of the network (e.g. centrality) are better recovered through lower burn probabilities (e.g., RW). Importantly, our recommendations lead to an average improvement of *KSD* of approximately 120% over the extant methods RN, EGO, RW, and SB across the four network characteristics. We next evaluate whether more information about a population structure may be used to further improve sampling performance.

4.2. Additional analyses and results

In some applications, we may have information about the population network prior to sampling. The next set of results help in selecting the most appropriate sampling method when such information is available. Specifically, we explore the interaction effects of the sampling method with two population characteristics, namely, population network type and degree density, as well as the interaction of sampling method with sample size. We also summarize here the results of several robustness checks that we undertook to assess the reliability and validity of our analyses and results.

4.2.1. Sampling method × network type (HK or WS)

For clustering coefficient, our general recommendation of a burn probability in the range of 40% to 80% turns out to be good for many different types of population networks, though it is not always uniquely best. In particular, FF50 generally performs well relative to the other subgraph sampling methods in both HK and WS networks. In no case that we list is its performance statistically worse than the other methods. We find similar results for the other local network characteristic, degree, for which we suggest the use of FF80 when nothing is known about the population. This is a robust recommendation across network types. EGO sampling method is strongly affected by network type. For WS, EGO is the *best* method (followed by FF80), but for HK, EGO is the *worst*

[6] We conjecture that a possible structural difference between HK network and the Facebook network is the extent of *community* or *modularity structure* i.e. *Q* (Clauset, Newman, and Moore, 2004). For the HK network, $Q = 0.245$ with the network being divided into 19 subgroups, whereas for Facebook $Q = 0.513$ with the network being divided into 31 subgroups. The stronger community structure is perhaps due to students being more likely to be connected on Facebook to others in the same dorm or the same class, for example, compared to their other Facebook friendships. An avenue for future research is to explore the effect of community structure on network sampling.

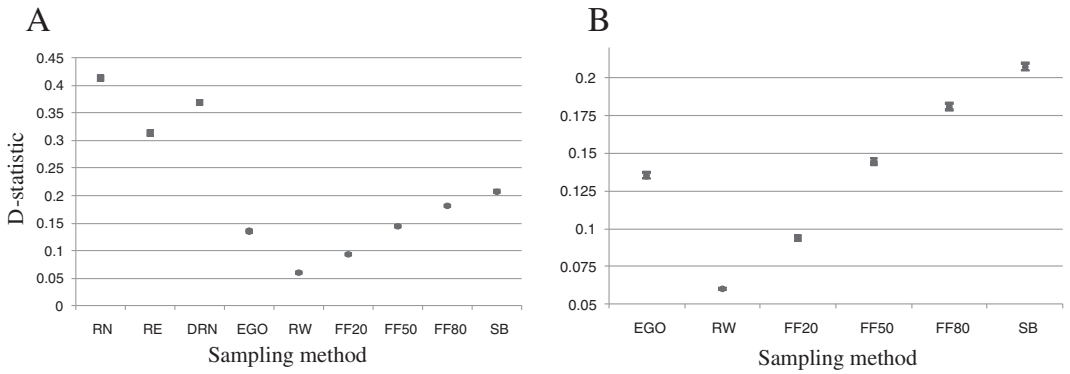

Fig. 5. Closeness D-statistic and its confidence interval by sampling method.

method (and FF50 the best). Hence, for recovering local network characteristics, FF50 and FF80 seem to provide the optimal balance.

When we consider the two global network characteristics, on average, across all simulated conditions, RW sampling was best in recovering the distribution of closeness and betweenness centrality, and knowledge of population network type does not lead to any further improvement over this general recommendation. For WS networks, however, the FF20 performs equally well as the general recommended method RW. Similar to the insights for local network characteristics, the performance of the recommended method(s) is much less affected by network type than the performance of other sampling approaches such as EGO or SB.

4.2.2. Sampling method × density

Changes in the density of the population network have the greatest effect on recovery of clustering coefficient (Fig. 1B). If the researcher implements our recommended method for clustering coefficient (FF50), the average change in *KSD* between high density and low-density networks was 0.26. When the network density increases, it adversely affects sampling performance. One way to address this problem is to increase the sample size to recover clustering coefficient in dense networks. For example, the *KSD* for FF50 for the smallest sample size case (100 nodes) in a low density network is 0.27 versus 0.49 in a high density network. Increasing the sample size to 2000 nodes in the high density case, however, results in a *KSD* of 0.35 using FF50. Hence, we may need large samples for obtaining the same level of recovery for clustering coefficient in networks that have high density levels, as compared to networks that have low density levels.

Table 7
Validation of the main recommendation in the three holdout networks.

		Clustering coefficient			Degree			Betweenness centrality			Closeness centrality		
		Mean *KSD*	Underperformance (%)		Mean *KSD*	Underperformance (%)		Mean *KSD*	Underperformance (%)		Mean *KSD*	Underperformance (%)	
HK/WS	Recommended	0.360	(FF50)		0.138	(FF80)		0.095	(RW)		0.060	(RW)	
	RN	0.930	158.08	*	0.225	63.10	*	0.466	390.14	*	0.413	588.43	*
	RW	0.392	8.78	*	0.170	23.70	*	0.095	0.00	NS	0.060	0.00	NS
	SB	0.400	11.10	*	0.151	9.39	*	0.272	185.72	*	0.207	245.48	*
	EGO	0.412	14.29	*	0.145	5.37	*	0.259	172.77	*	0.135	125.37	*
PA/CC	Recommended	0.246	(FF50)		0.154	(FF80)		0.093	(RW)		0.054	(RW)	
	RN	0.854	247.92	*	0.219	42.14	*	0.422	351.33	*	0.385	615.03	*
	RW	0.236	−3.81	NS	0.164	6.28	*	0.093	0.00	NS	0.054	0.00	NS
	SB	0.293	19.33	*	0.158	2.47	NS	0.256	174.00	*	0.197	265.19	*
	EGO	0.285	16.19	*	0.145	−6.15	*	0.234	150.21	*	0.124	129.11	*
Facebook	Recommended	0.260	(FF50)		0.084	(FF80)		0.163	(RW)		0.059	(RW)	
	RN	0.801	208.39	*	0.253	202.05	*	0.350	114.50	*	0.369	523.75	*
	RW	0.264	1.62	NS	0.119	42.00	*	0.163	0.00	NS	0.059	0.00	NS
	SB	0.316	21.64	*	0.078	−7.11	NS	0.114	−29.80	*	0.099	66.52	*
	EGO	0.405	55.70	*	0.088	5.50	NS	0.125	−23.18	*	0.109	83.55	*
Coauthorship	Recommended	0.125	(FF50)		0.175	(FF80)		0.112	(RW)		0.124	(RW)	
	RN	0.881	605.55	*	0.451	157.06	*	0.615	448.61	*	0.540	336.46	*
	RW	0.141	12.79	NS	0.261	49.13	*	0.112	0.00	NS	0.124	0.00	NS
	SB	0.173	38.08	*	0.173	−1.43	NS	0.126	12.50	*	0.169	36.65	*
	EGO	0.237	89.91	*	0.198	12.80	*	0.180	60.58	*	0.173	40.06	*

Notes: Reported figures are the underperformance measures $100 \times (\frac{\text{KSD of recommended method}}{\text{KSD of benchmark}} - 1)$. An '*' indicates that difference in performance between benchmark sampling approach and recommended approach is significant, and NS indicates not significant, for $\alpha = 5\%$.

4.2.3. Higher-order interactions: knowledge about both density and network type

Although modest improvements over our general recommendations may be obtained with knowledge about density and/or network type, is it possible to improve sampling further if the researcher has knowledge about *both* population density and network type? Higher-order interactions are most relevant for recovery of clustering coefficient and betweenness centrality (Table 6) and least relevant for recovery of degree distribution. We can see an interesting pattern (Fig. 6) when we examine the three-way interaction effects of sampling method, network type, and density level for clustering coefficient. We observe that for low-density HK networks, the EGO and RW methods deliver better or comparable *KSD*, as FF50 or FF80. But, for high-density HK networks, the general recommendation (FF50) remains the best. However, for the WS network, we see diverging patterns and our general recommendation (FF50) may be improved upon; for low density WS networks, higher burn probabilities (e.g., SB) are preferable whereas for high-density WS networks, lower burn probabilities (FF20) achieve the best performance.

4.2.4. Sampling method × sample size

A small sample size is often preferred in marketing applications for cost reasons. The previous results suggest that for high-density networks, larger samples sizes are preferable. In general, however, if the researcher implements our recommended method for clustering (FF50), degree (FF80), and centrality (RW), the average change in *KSD* between large samples (≥700) and small (≤300) samples is only 0.029.

4.2.5. Cross-entropy versus KSD

If we compare the results from the cross entropy metric with *KSD*, the recommendations are identical for clustering coefficient (FF50), and for betweenness and closeness centrality (RW). In the case of recovery of degree distribution, EGO performed slightly better than our recommended method (FF80) based on *KSD*. In fact, EGO performed slightly better overall under cross-entropy measure than it did with the *KSD*. For conditions with higher-order interactions, there were no major systematic differences in results between these two measures of sample quality; however, we observed some differences in the rank-orders of performance of the sampling methods in some particular cases.

In sum, a detailed examination of the performance of our main sampling recommendations indicates that they are generally robust across the simulated conditions. However, we do find that with EGO sampling researchers should be aware of the possibility of highly variable performance depending on the characteristics of the population network. Further, we find that for sample sizes of up to 20% of the population, the performance of the recommended sampling methods is affected only modestly for all four network characteristics across the simulated conditions. Sample size considerations appear to be most important in networks of higher density.

4.3. Network sampling and process recovery

In the previous sections we discussed results pertaining to recovery of network structure through subgraphs. Here we investigate whether improved recovery of network structure in a sample also improves recovery of a population process obtained by observing that process in the subgraph.

We consider two types of process studies of interest to marketers: (1) Studies in which a marketer wishes to understand how a social process (e.g. new product diffusion) would take place in a population based on observing how that process plays out in a sampled network. We consider node-level adoptions of a new product (a local characteristic) and the influence of an adopting node on downstream adoptions (a global characteristic) as the two outcome measures of interest. (2) Studies in which a marketer is interested in spotting 'hubs'; that is, finding nodes with a high degree compared to other nodes. Identifying hubs may be relevant for testing their influence on adoptions that take place over the network (Peres, 2014). We note that if a complete list of nodes (sampling frame) is available, then a random sample of just the nodes may suffice for spotting hubs (e.g. in a survey, ask

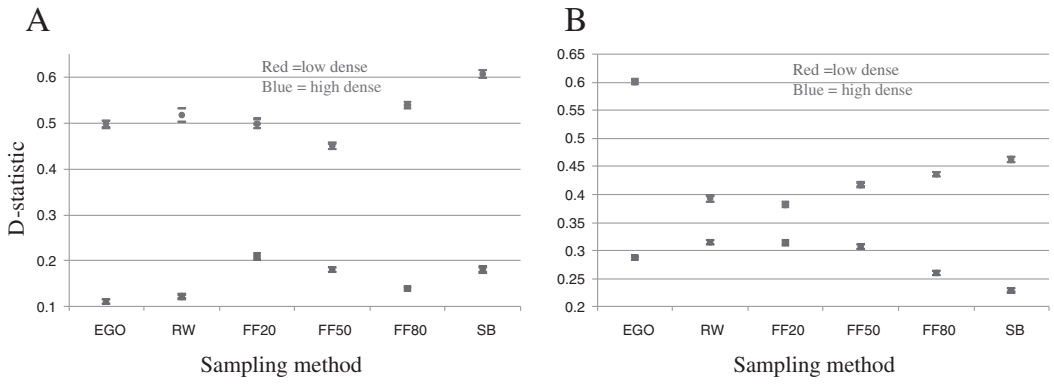

Fig. 6. Clustering coefficient D-statistic and its confidence interval by sampling method & network type & density level (low dense: avg. degree = 4; high dense: avg. degree = 16) for HK (left) and WS (right).

the sampled nodes how many friends they have). Likewise, we could use a simple random sample of just the nodes to determine node-level adoption (e.g. in a survey, ask the sampled nodes whether they adopted) as described in Section 2.3. In this section, however, we implicitly assume an application context where the population network is not fully observed by the marketer and network sampling is required to both understand the structure of the network and to understand the processes that take place over that network.

4.4. Network-based diffusion process

We used the threshold model of interpersonal influence analyzed by Watts and Dodds (2007) to simulate a diffusion process in a social network (see Appendix III for details). We simulate in the sampled subgraph the same diffusion process as in the population. We consider two diffusion process metrics:

(1) *Adoption proportion* is the proportion of individuals who adopted. We expect that to correctly predict adoption proportions, it is important to recover the local structural characteristics of networks. Accordingly, we hypothesize that our sampling recommendations based on recovery of clustering coefficient (FF80) and degree (FF50) will result in a more accurate measure of adoption proportions.
(2) *Downstream influence* is a node-level influence measure to capture how the adoption of the product by node *i* influences the adoption of all other nodes that are connected directly or indirectly to node *i* but adopt after node *i*. We expect that to correctly predict downstream influence, it is important to recover the global structural characteristics of networks. Accordingly, we hypothesize that our sampling recommendation (RW) based on recovery of betweenness and closeness centrality will result in more accurate measures of the process measure of downstream influence.

4.4.1. Results for process inference

Table 8 summarizes the results from our process simulations. We find that the subgraph structure makes a substantial difference in accurately recovering the outcomes of processes that take place in a population. We measure the accuracy by comparing the results from our recommended methods versus the following sampling methods: RN, EGO, RW, and SB. We find that the recommended sampling methods based on degree and clustering coefficient (FF50 and FF80) produced gains of 75.84% and 75.37% over RN, 41.73% and 40.62% over EGO, 39.66% and 38.51% over RW, and 5.42% and 3.62% over SB, respectively, in recovering the population values of aggregate adoption proportions. Note that the poor performance of RN occurs because we measure adoption proportion from a process that takes place in the sampled subgraph, rather than in the population network. If a diffusion process has already taken place in the population network, then we can obtain an unbiased estimate of the adoption proportion from a simple random sample of just the nodes, if a complete list of nodes is available and if we can observe their adoption behavior without error.

Similarly, the recommended sampling method based on centrality measures (RW) produced overall gains of 83.36% over RN, 56.28% over EGO, and 43.25% over SB, in terms of the *KSD* values of the influence distribution. Based on our analysis, we conclude that developing an appropriate subgraph of a population network is important for understanding the evolution of systemic processes, such as adoption and influence that could occur in that population network.

4.4.2. Results for hub spotting

Social hubs are likely to be present in scale-free networks, such as the HK networks, where the degree distribution is power law. To identify social hubs we conjecture that for a given sample size, sampling methods that go deeper into the network (e.g., RW) perform better. The reason is that in order to assess whether a node is a possible hub, one needs to add nodes to the sample that are more than a few steps away from already sampled nodes, to assess a node's position in the network relative to other nodes.

To spot hubs, we count the number of nodes that are one, two, and three standard deviations away from the mean degree, and we divide this number by the size of the network. This measure reflects the probability of randomly drawing a hub from a set of nodes. We compute the squared difference between the relative number in a sample and the relative number in the population as a measure for sample quality. We compared the results of our recommended method (RW) to the traditional methods RN, EGO, and SB. The results are fairly consistent (see Table 8). We find that on average RW performs better relative to RN, EGO, and SB, which have underperformance measures of 319.03%, 21.60%, and 95.02%, respectively.

Table 8
Results from process recovery simulations.

	Adoption proportion		Downstream influence	Hub spotting
Recommended	(FF50)	(FF80)	(RW)	(RW)
RN	75.84	75.37	83.36	319.03
RW	39.66	38.51	0.00	0.00
SB	5.42	3.62	43.25	95.02
EGO	41.73	40.62	56.28	21.60

The figures reported in the table are the underperformance measures (in %) with respect to our recommended method.

In sum, improved recovery of network structure enhances the analyses of processes that take place on the network. The sampling method that performs the best depends on whether the focus is on local or global aspects of the population network. Our initial process simulation results here are suggestive, and a more comprehensive study involving multiple representative network-based process mechanisms is required to fully address the generalizability of our structure-based findings for network process inference.

5. Conclusions, limitations, and future research

5.1. Conclusions

Marketing has had a long-standing interest in social networks and in the interactions and exchanges between individuals that take place in those networks. An emerging opportunity for marketers is the potential for exploiting social-network-based marketing strategies; however, it is often prohibitively expensive to do careful studies of the effects of a marketing effort on a social network when the network is large. Our findings indicate that generating good subgraphs of population networks through sampling may be a cost-effective way for marketers to obtain more accurate network-level outcome measures about individual and group influence that result from marketing efforts. Representative subgraphs also enable marketers to design and conduct field experiments or simulations to predict how local micro flows (e.g. emanating from seeded nodes) generate systemic flows or cascades.

Deeper knowledge about the social network of their target customers could be particularly useful for marketers who are considering offering free samples (e.g., free storage for referring a friend to Dropbox; free Uber rides for inviting a friend to signup) or referral programs (e.g., cash for referring a friend to Paypal or Airbnb). Paypal's refer-a-friend bonus (which gave $10 bonus to both referrer and referee) during the early stages of its launch was partly responsible for the increase its user base from just 12,000 at the beginning of the year 2000 to 2.2 million six months later.[7] Dropbox increased its user base from 100,000 in Sept. 2008 to 4 million users in just 15 months, of which 35% of the growth is attributed to a referral program which offered free additional storage space to current customers who referred friends who signed up for the service.[8] These are two examples of the successful use of social connections to drive business outcomes. In both cases, the "seeding" of the process was done using either judgmental or convenience samples, and without a clear understanding of the social structures connecting the members of their target populations. However, based on the insights from our study, for example, the campaigns could have been done differently. We conjecture that had the companies determined the nature of the social network connecting members of their target populations via sampling, they potentially could have optimized their programs further through small scale experiments on the selected sample.

We summarize below our main findings and contributions about how to sample social networks:

- Marketing studies in which the focus is on understanding local influence (e.g., influence of friends, family or workgroups) should use medium burn-rate forest-fire sampling in an attempt to get improved recovery of clustering coefficient and degree. Marketing studies that focus on network-level (or distant) influence, should use low-burn forest-fire sampling (e.g. RW) focused on recovering node centrality. These are simple and useful guidelines that emerged from our study (Section 2.3; Table 2).

The intuition for the above guidelines can be summarized as follows: For a given sample size, a lower burn rate forces the subgraph sampling procedure to reach nodes farther away from the starting nodes (i.e., the sampling penetrates deeper into the network), which enables improved measurement of global network properties. Conversely, given the same sample size, a higher burn rate keeps the sampled nodes closer to the starting nodes, because a large proportion of the close neighbors of the starting nodes will now be included in the sample, which should enable improved measurement of local network properties.

- Traditional sampling methods such as RN, RE, and DRN sampling perform poorly in recovering population network structure, as compared to subgraph sampling methods, and should be avoided in sampling social networks when the focus of the study is not just about nodal characteristics, but extends to network effects. In fact, just studying nodal characteristics in a network defeats the purpose of studying social networks, because the important drivers of node behaviors in networks are the connections between the nodes, and not just the intrinsic aspects of the nodes themselves.
- By following our recommended sampling strategies, a marketer could expect an overall average improvement of approximately 120% in *KSD* over the extant methods RN, EGO, RW, and SB across the four network characteristics. Importantly, the recommended sampling approaches are not strongly affected by different aspects of the population network structures considered in our simulation study, unlike other sampling approaches that are sensitive to population network structure (e.g. EGO).
- Our agent-based process simulation study suggests that a more accurate representation of social network structure in the sampled subgraph allows for more accurate inferences of network processes, by studying the processes on the sampled subgraph instead of the population graph. This should enable better decision making using sampled networks, which offers the managerial justification for following our recommendations.
- Overall, changes in sample size had a modest effect on both absolute and relative performances of the sampling methods. This result is reassuring and suggests that even modest-size samples will offer benefits in terms of improved accuracy of process and outcome measures associated with social networks. However, having a larger sample is more important in cases where the

[7] From 10-K filing for year ending December 2001.
[8] From a presentation made by Drew Houston, founder, on April 23, 2010 at the Startup Lessons Learned Conference, San Francisco.

network is dense. In other cases, samples that are at around 1–3% of size of the population network may be sufficient when considering both the costs and accuracy of the results obtained from samples. For improving recovery of the population network characteristics, it is important to ensure that the correct sampling method is used, before considering larger sample sizes.
- The concept of a "mini-population" that we introduced offers a way to construct benchmark networks to compute the upper bound on the performance we can expect from the best performing sampling method(s) with a given sample size. This approach is useful for testing proposed new estimators of network characteristics where the estimates are obtained from sampled subgraphs generated from either existing or new sampling methods.

5.2. Limitations and future research

Our recommendations should be sufficient for practical marketing problems when the objective of network sampling is to understand either a (1) single network characteristic, or (2) local aspects of a network (e.g. peer influence), or (3) broader aspects of the network (e.g. speed of diffusion). However, further research is needed for developing specific recommendations for situations in which it is difficult to directly translate a marketing problem into one of these three categories. Specifically, future research could explore whether burn probabilities should be adjusted based also on demographic information obtained at a node (e.g., decrease burn rate for households with children). Likewise, recovery of other network characteristics such as degree-degree correlation and community structure may be worth exploring to the extent they are relevant for marketing tasks.

An understanding of the structure of large social networks via sampling is a logical first step before we can gain an accurate understanding of how various marketing processes would evolve in a network. Through a simulation, we explored a couple of processes on sampled networks (e.g. downstream influence), but we need a more comprehensive study, including an empirical application, to understand the conditions that favor the generalizability of processes (and associated outcomes) in the sampled network to the population network. Fortunately, more data of consumer behaviors at the individual level are becoming available, and future research could explore managerial consequences of network sampling and inferring network processes on empirical networks. For example, in the case of new product adoptions, we need information on the exact times (or, less restrictively, the exact sequence) in which a large group of connected adopters adopts a new product or service (see, for example, Yoganarasimhan, 2012; Shaikh, Rangaswamy, and Balakrishnan, 2010). In such studies, it should be feasible to also explore the efficacy of our sampling recommendations for recovering process measures based on a broad range of social influence mechanisms, such as reciprocity and compliance (see, for example, Kelman, 1958), instead of just the imitation-based diffusion process we explored in our simulations.

Another important direction for future research is to evaluate implementation issues related to subgraph sampling, particularly the field work challenges in the context of network sampling, such as (survey) instrument development, data collection, and respondent participation (see, for example, Marsden, 2011).

To conclude, our results provide an initial set of important new insights regarding sampling of large scale social networks for marketing applications. The implication for practice is clear: By judiciously selecting a sampling method following our recommendations, it is possible to improve recovery of population network characteristics and associated network process metrics even in samples of reasonable size.

Appendix I. Computational aspects of the network characteristics used in the study

I.1. Degree of a node

The *degree* of a node in a network is the number of edges connected to that node. Let p_k denote the degree distribution of the graph, which is the probability that a node chosen uniformly at random has degree k. For instance, scale-free networks have a power-law degree distribution $p_k \sim k^{-\alpha}$ where α is a positive constant, whereas small world networks have (approximately) a Poisson degree distribution.

I.2. Clustering

To quantify clustering coefficient, we adopt the definition proposed by Watts and Strogatz (1998), and define a clustering coefficient for each node that has at least two neighboring nodes as

$$CL_j = \frac{\text{number of edges between neighbors of nodes } j}{k_j\left(k_j - 1\right)/2} \tag{A1}$$

where k_j is the degree of the j-th node.

I.3. Centrality

The path length between two nodes is defined as the number of edges along the shortest path connecting them. We study two versions of node-based centrality measures derived from the path length. First, *betweenness centrality* captures the idea that a well-placed node that lies on the shortest paths between other nodes has the potential to control communication in the network and command attention (Freeman, 1979). The betweenness centrality of a node j is defined as the sum of the fraction of all-pair shortest paths that pass through j:

$$BC_j = \sum\nolimits_{s,t \in V, s,t \neq j, \sigma(s,t)>0} \frac{\sigma(s,t|j)}{\sigma(s,t)} \tag{A2}$$

where V is the set of nodes in the network, $\sigma(s, t)$ is the number of shortest paths between nodes s and t, and $\sigma(s, t|j)$ is the number of shortest paths between nodes s and t that pass through node j.

Second, *closenesses centrality* of a node j is the inverse of the average shortest number of hops in a network that connects that node to all other nodes in the network. We compute closeness centrality of node j only in the connected part G_j of the network containing node j and then multiply that value by the ratio between the size of the G_j and G to ensure that the computed value properly reflects the effect of connected parts of different sizes. Specifically,

$$CC_j = \frac{N_j - 1}{N-1} \frac{N_j - 1}{\sum_{s \in V(G_j), s \neq j} l(j,s)} \tag{A3}$$

where $V(G_j)$ is the set of nodes in the connected part G_j containing node j, $l(j, s)$ is the shortest path length between nodes j and s, N_j is number of nodes in G_j, and N is number of nodes in G. This is the implementation adopted in the NetworkX package (http://networkx.lanl.gov/).[9]

Appendix II. Description of population graph generation

II.1. Watts–Strogatz (WS) network (Watts and Strogatz, 1998)

We start with a ring of n nodes, and then link each node to its K nearest neighbors. Subsequently, a "random rewiring" process is performed. Specifically, each edge (v_i, v_j) is replaced with probability p with a new edge (v_i, v_w), with v_w chosen uniformly random from the existing nodes with duplicate edges forbidden. It can be shown that $L \sim n/2\,K$ and $C \sim 3/4$ as $p \to 0$, while $L \approx L_{\text{random}} \sim \ln(n)/\ln(K)$ and $C \approx C_{\text{random}} \sim K/n$ as $p \to 1$, where L and L_{random} are average path length and C and C_{random} are the average clustering coefficient of the WS graph and Erdös–Rényi random graph (Erdös and Rényi, 1959), respectively. The WS network is very flexible in that there is a broad interval of p over which $L(p)$ is almost as small as L_{random} yet $C(p)$ far exceeds C_{random}. The degree distribution of the WS network can be written as (Barrat and Weigt, 2004):

$$p_k = \sum_{d=0}^{\min(k-K/2, K/2)} \binom{K/2}{d} (1-p)^d p^{K/2-d} \frac{(Kp/2)^{k-K/2-d}}{(k-K/2-d)!} \exp(-pK/2), k \geq K/2. \tag{A4}$$

The shape of the degree distribution is similar to that of a random graph and has a pronounced peak at $k = K$ and decays exponentially for large $|k - K|$. We have set the rewiring probability p to be 0.3 and varied the initial neighborhood size K to be 4, 8, 10, and 16 to obtain the four WS population networks with different levels of density.

II.2. Holme–Kim (HK) network (Holme and Kim, 2002)

The graph generation starts with m_0 nodes and no edges and then a node v with m edges is added one at a time. These m edges are constructed by linking v with an existing node w with a probability proportional to the degree of w:

$$P_w = \frac{k_w}{\sum_{u \in V} k_u} \tag{A5}$$

where k_w is the degree of node w in the current graph with the node set V. If edge (v, w) was added in the previous step, then with probability p one more edge from v to a randomly chosen neighbor of w is added (thus forming a triangle). It was shown that the HK network has (1) a power law degree distribution $p_k \sim k^{-a}$ with $a \approx 3$; (2) a finite clustering coefficient for $N \to \infty$ that is a function of p; and (3) a path length L which increases logarithmically with N. In our study, we have set the triangle formation probability p to be 0.4 and varied the parameter m to be 4, 8, 10, and 16 to obtain the four HK population networks with different levels of density.

[9] Last accessed: August 2015.

II.3 Rewired connected caveman (CC) network (Watts, 1999)

The CC network is constructed using a three-step procedure. First, a caveman network is created consisting of isolated fully connected subgraphs (caves). In our study, the size of the subgraph is the same as the pre-specified average degree. Then one link in each subgraph is used to link to a node in another subgraph such that all subgraphs are arranged in a ring structure. Now we have a connected caveman model, which is a connected network with the largest possible clustering coefficient given a certain average degree. This network, however, has large average path length. We then perform the random rewiring procedure similar to that of the WS network (same rewiring probability of 0.30) to form the rewired connected caveman network that has a large clustering coefficient and short path length.

II.4. Barabasi–Albert (BA) network (Barabási and Albert, 1999)

The BA network was proposed to produce the empirically observed power-law distribution of the node degree. The model utilizes a simple process of growing a network by attaching new nodes each with m edges that are preferentially attached to existing nodes with high degree, where m determines the average degree or density of the network. The model is also referred to as the preferential attachment model. The resulting network typically exhibits power-law degree distribution (due to preferential attachment) and short path length (due to random attachment) but low clustering coefficient.

Appendix III. Network-based diffusion process

We used the threshold model of interpersonal influence analyzed by Watts and Dodds (2007) to simulate a diffusion process in a social network. Each individual in this network makes a binary decision A (e.g., not adopt) or B (e.g., adopt) using a threshold rule, which posits that individuals will switch from A to B only when sufficiently many others have adopted so that the perceived benefits of B outweighs A. Specifically, in our diffusion process, all individuals start in the initial state A. A small percentage of randomly selected nodes (seed nodes) are then set to state B. After this initialization stage, at each time period, each node i with state A makes a binary decision whether or not to switch to state B according to the following threshold rule: Switch from A to B if the percentage of the neighbors in state B is greater than or equal to a node-specific threshold φ_i. The diffusion process stops when a specified saturation percentage of nodes in state B has been reached in the network, or no more switching is possible. For purposes of illustration, we stopped the simulation at the end of time 6, by which time about 83% of the adoptions in WS networks and 94% of adoptions in the HK network had already taken place.

III.1. Process description and parameters

Our results are based on the following settings: We set the number of seed nodes at initialization to 30 (or 0.3% of the 10,000 nodes) and φ_i follows a uniform distribution U[0, 0.2] (Watts and Dodds, 2007). We use two population models and one density level for this illustration: the HK — high-density and WS — high-density models, as summarized in Table 4 of the paper. We study the process correspondence in subgraphs that are 10% of the population network where the subgraphs were obtained from the different sampling methods. We simulate in the sampled subgraph the same diffusion process as in the population (i.e., the number of seed nodes is now 3, which is also 0.3% of the network size). We consider two diffusion process metrics:

(1) *Adoption proportion* is the proportion of individuals who adopted (switched from state A to B) at time t. We assess sampling performance as the absolute difference between the sample and population adoption proportions.
(2) *Downstream influence* is a node-level influence measure to capture how the adoption of the product by node i influences the adoption of all other nodes that are connected directly or indirectly to node i but adopt after node i. Specifically, for each individual i, we obtain a set of followers $F_i = \{j\}$, where a node j is deemed a follower of i if a path $(k_1, k_2 \ldots, k_m)$ can be found so that $i = k_1, j = k_m, t(k_d) \leq t(k_d + 1)$ and k_d is connected to $k_d + 1$. Here $t(k)$ denotes the time when node k adopted and is only defined for nodes that have adopted by a given time t. We define the influence measure for each node i at time t as the relative size of the set of followers $|F_i|/N$ up to time t, where N is the size of the network. Nodes have a greater influence if they have a larger proportion of followers who adopt. We compare the sample distribution of the influence measure to its population distribution using the *KSD* statistic, as we did in the main simulations reported earlier.

References

Albert, R., & Barabási, A. -L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.
An, W. (2011). Models and method to identify peer effects. In J. Scott, & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 514–532). London (UK): Sage Publications.
Aral, S. (2015). Networked experiments: A review of methods and innovations. In Y. Bramoulle, A. Galeotti, & B. Rogers (Eds.), *The Oxford handbook on the economics of networks*. Oxford University Press (in press).
Barabási, A. -L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.
Barrat, A., & Weigt, M. (2004). On the properties of small-world network models. *European Physical Journal B: Condensed Matter and Complex Systems*, *13*(3), 547–560.
Capobianco, M., & Frank, O. (1982). Comparison of statistical graph-size estimators. *Journal of Statistical Planning and Inference*, *6*(1), 87–97.

Chen, X., Chen, Y., & Xiao, P. (2013). The impact of sampling and network topology on the estimation of social inter-correlations. *Journal of Marketing Research*, *L*, 95–110.
Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*, 06611.
Doreian, P., & Woodard, K. L. (1992). Fixed list versus snowball selection of social networks. *Social Science Research*, *21*, 216–233.
Dover, Y., Goldenberg, J., & Shapira, D. (2012). Network traces on penetration: Uncovering degree distribution from adoption data. *Marketing Science*, *31*(4), 689–712.
Erdös, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, *6*, 290–297.
Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, *1*, 91–101.
Frank, O. (1979). Estimating of population totals by use of snowball sampling. In P. W. Holland, & S. Leinhardt (Eds.), *Perspectives on social network research* (pp. 319–347). New York: Academic Press.
Frank, O. (2011). Survey sampling in networks. In J. Scott, & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 389–403). London (UK): Sage Publications.
Frank, O. (2012). Social network analysis, estimation and sampling. In R. A. Meyers (Ed.), *Computational complexity* (pp. 2845–2863). New York: Springer.
Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*, 215–239.
Goldenberg, J., Han, S., Lehmann, D., & Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of Marketing*, *73*(March), 1–13.
Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, *32*, 148–170.
Granovetter, M. (1976). Network sampling: Some first steps. *The American Journal of Sociology*, *81*(6), 1287–1302 May.
Handcock, M. S., & Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of Applied Statistics*, *4*(1), 5–25.
Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, *75*, 55–71.
Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical Reviews Letters E*, *65*, 026107.
Iyengar, R., van den Bulte, C., & Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*(2), 195–212.
Kelman, H. C. (1958). Compliance, identification, and internationalization — Three processes of attitude change. *Conflict Resolution*, *II*(1), 51–60.
Klovdahl, A. S. (1977). Social networks in an urban area: First Canberra study. *Australian and New Zealand Journal of Sociology*, *13*, 169–175.
Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models.* LLC: Springer Science + Business Media.
Lakhina, A., Byers, J. W., Crovella, M., & Xie, P. (2003). Sampling biases in IP topology measurements. *Proceedings of the 22nd annual joint conference of the IEEE computer and communications societies* (pp. 332–341).
Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 631–636).
Levinea, S. S., & Kurzbanb, R. (2006). Explaining clustering in social networks: Towards an evolutionary theory of cascading benefits. *Managerial and Decision Economics*, *27*, 173–187.
Libai, B., Muller, E., & Peres, R. (2013). Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of Marketing Research*, *50*(April), 161–176.
Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM), New Orleans, LA* (pp. 556–559).
Lindgren, B. W. (1993). *Statistical theory.* Boca Raton: Chapman & Hall/CRC.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: Wiley-Interscience.
Marsden, P. V. (2011). Survey methods for network data. In J. Scott, & P. J. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370–388). London (UK): Sage Publications.
Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 5th ACM/USENIX internet measurement conference (IMC'07), San Diego, CA* (pp. 29–42).
Nair, H. S., Manchanda, P., & Bhatia, T. (2010). Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *Journal of Marketing Research*, *47*(5), 883–895.
Narayan, V., Rao, V. R., & Saunders, C. (2011). How peer influence affects attribute preferences: A Bayesian updating mechanism. *Marketing Science*, *30*(2), 368–384.
Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physics Review Letters E*, *64*(016131), 1–8.
Peres, R. (2014). The impact of network characteristics on the diffusion of innovations. *Physica A*, *402*, 330–343.
Reingen, P. H., & Kernan, J. B. (1986). Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research*, *23*(4), 370–378.
Salganik, M. T., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *34*, 193–239.
Scosyrev, E. (2014). Estimation of population mean under unequal probability sampling with unknown selection probabilities. *American Journal of Theoretical and Applied Statistics*, *3*(3), 65–72.
Shaikh, N. I., Rangaswamy, A., & Balakrishnan, A. (2010). Modeling the diffusion of innovations through small-world networks. *(Working paper, Smeal College of Business, Penn State University).*
Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*, 268–276.
Thompson, S. K. (2006). Targeted random walk designs. *Survey Methodology*, *32*(1), 11–24.
Thompson, S. K. (2012). *Sampling* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
Thompson, S. K., & Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, *26*, 87–98.
Toubia, O., & Stephen, A. T. (2013, May–June). Intrinsic vs. image-related utility in social media: Why do people contribute content to Twitter. *Marketing Science*, *32*(3), 368–392.
Trusov, M., & Rand, W. (2013). Improving prelaunch diffusion forecasts: Using synthetic networks as simulated priors. *Journal of Marketing Research*, *50*(6), 675–690.
Trusov, M., Bodapati, A., & Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of Marketing Research*, *47*(August), 643–658.
Valente, T. W., Watkins, S., Jato, M. N., Van der Straten, A., & Tsitsol, L. M. (1997). Social network associations with contraceptive use among Cameroonian women in voluntary associations. *Social Science and Medicine*, *45*, 677–678.
Van den Bulte, C., & Wuyts, S. (2007). *Social networks and marketing.* Cambridge, MA: Marketing Science Institute.
Wasserman, S., & Faust, K. (1995). *Social network analysis.* UK: Cambridge University Press.
Watts, D. J. (1999). *Small worlds: The dynamics of networks between order and randomness.* Princeton, N.J.: Princeton University Press.
Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*, 441–458.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*, 440–442.
Yoganarasimhan, H. (2012). Impact of social network structure on content propagation: A study using YouTube data. *Quantitative Marketing and Economics*, *10*(1), 111–150.