Applied Network Science

RESEARCH

Open Access

Inclusive random sampling in graphs and networks

Yitzchak Novick[1,2*] and Amotz Bar-Noy[3]

*Correspondence:
ynovick@gradcenter.cuny.edu;
yitzchak.novick@touro.edu

[1] CUNY Graduate Center, New York, NY 10016, USA
[2] Touro University, New York, NY 10018, USA
[3] Brooklyn College, Brooklyn, NY 11210, USA

Abstract

It is often of interest to sample vertices from a graph with a bias towards higher-degree vertices. One well-known method, which we call random neighbor or RN, involves taking a vertex at random and exchanging it for one of its neighbors. Loosely inspired by the friendship paradox, the method is predicated on the fact that the expected degree of the neighbor is greater than or equal to the expected degree of the initial vertex. Another method that is actually perfectly analogous to the friendship paradox is random edge, or RE, where an edge is sampled at random, and then one of the two endpoint vertices is selected at random. Obviously, random sampling is only required when full knowledge of the graph is unattainable. But, while it is true in most cases that knowledge of all vertices' degrees cannot be obtained, it is often trivial to learn the degree of specific vertices that have already been isolated. In light of this, we suggest a tweak to both RN and RE, inclusive random sampling. In inclusive random neighbor (IRN) the initial vertex and the selected neighbor are considered, in inclusive random edge (IRE) the two endpoint vertices are, and in both cases, we learn the degree of each and select the vertex of higher degree. This paper explores inclusive random sampling through theoretical analysis and experimentation. We establish meaningful bounds on IRN and IRE's performances, in particular in comparison to each other and to their exclusive counterparts. Our analyses highlight differences of the original, exclusive versions as well. The results provide practical insight for strategizing a random sampling method, and also highlight graph characteristics that impact the question of which methods will perform strongly in which graphs.

Keywords: Inclusive random sampling, Random neighbor, Random edge

Introduction

Finding high-degree vertices in a graph is an important goal in many endeavors. A few examples include network immunization (Cohen et al. 2003), early detection of network phenomena (Christakis and Fowler 2010), and locating network influencers (Malliaros et al. 2016) among many others. Naïvely sampling a random vertex, a method we call RV, will return a vertex whose expected degree is the mean degree of a graph. Because total knowledge of the graph is usually impossible to obtain, there is typically no way to target high-degree vertices directly. One well-known sampling method that is effective for finding high-degree vertices is random neighbor, or RN (Cohen et al. 2003) (see also Momeni and Rabbat 2018). Like RV, a vertex is sampled at random, but then it is

exchanged for one of its neighbors. The expected degree of this selected neighbor is higher than that of the first vertex, in concert with the message of Scott Feld's friendship paradox (Feld 1991) that, on average, friends have a mean-degree greater than or equal to individuals. A lesser-known method is random edge (RE) (Leskovec and Faloutsos 2006; Pal et al. 2019), which also returns a vertex whose expected degree is greater than or equal to the mean degree of the graph. In RE, an edge is sampled at random from the edges of the graph and one of the two endpoint vertices is then selected.

Our research proposes a novel tweak to both of these methods. While is it true that learning the degree of all vertices in a graph is typically not possible, learning the degrees of a few selected vertices is often not only possible, but trivial. In both RN and RE, two vertices are isolated before one is ultimately selected. If we learn the degrees of the two vertices, we can select the one of higher degree, thereby correcting for specific limitations in the sampling methods. We call these methods "inclusive random sampling", specifically "inclusive random neighbor" or IRN, and "inclusive random edge" or IRE.

This paper extends our previously published introduction of this topic (Novick and Bar-Noy 2020). In this paper, we offer an extensive exploration of all four methods under discussion, RN, RE, IRN, and IRE. We compare and contrast all of these methods using both theoretical and experimental analyses and establish important bounds on some of the main comparisons. We include a number of results that are either new, or were omitted from the previous paper for brevity, such as the upper bound on $\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]}$, and an experimental analysis of the role of the power-law exponent in predicting the strengths of the methods. A number of new equations are included and the full proofs of the unbounded nature of the $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$ and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ ratios are presented as well. This full exploration of inclusive random sampling elucidates many of the theoretical aspects of the sampling methods and suggests practical ideas for strategizing a sampling approach when certain graph characteristics are known.

Background

This section summarizes the RN and RE sampling methods and presents some of the existing research which is fundamental to our findings.

RN

The random neighbor sampling method was introduced by Cohen et al. (2003). The suggestion is that a neighbor of a vertex will have the higher expected degree, so an initially sampled vertex is exchanged for one of its neighbors that is selected at random. The superiority of the sampling method is often attributed to Scott Feld's friendship paradox (Feld 1991), the network phenomenon that the collection of "friends" in a network have a mean degree greater than or equal to the mean degree of the graph. This explanation is erroneous though, and this is demonstrated by Kumar et al. (2018) with a simple counterexample. Construct a graph comprised of a clique of four vertices, and an additional two vertices connected to each other by a single edge, see Fig. 1. There is a variance of degree in the graph, so the FP holds. Yet, by symmetry, we know that the expected degree of a vertex returned by RN is equal to the expected degree of a vertex returned by RV, which we denote as $\mathbb{E}[RN] = \mathbb{E}[RV]$. It is always true though that $\mathbb{E}[RN] \geq \mathbb{E}[RV]$, and furthermore that $\mathbb{E}[RN] > \mathbb{E}[RV]$ in all graphs with at least one edge that connects

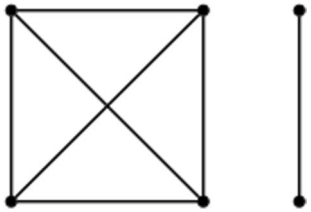

Fig. 1 A graph where the FP holds, yet RN reduces to RV

two vertices of different degree (Kumar et al. 2018; Novick and Bar-Noy 2022; Strogatz 2012).

We can calculate the expected degree of a vertex sampled by RN as

$$\mathbb{E}[RN] = \frac{1}{n} \sum_{v \in V} \sum_{u \in N(v)} \frac{d_u}{d_v} \tag{1}$$

where V is the set of vertices in the graph, n is the number of vertices in V, d_v and d_u are the degrees of v and u respectively, and $N(v)$ is the set of neighbors of vertex v.

It is worth noting that the contribution of every edge $e(u, v)$ to the outer summation is $\frac{d_u}{d_v} + \frac{d_v}{d_u}$ and therefore $\mathbb{E}[RN]$ can also be expressed as a summation over E, the set of edges in the graph.

$$\mathbb{E}[RN] = \frac{1}{n} \sum_{e(u,v) \in E} \left(\frac{d_u}{d_v} + \frac{d_v}{d_u} \right) \tag{2}$$

RE

In (Kumar et al. 2018), Kumar et al. distinguish between two types of "means of neighbor's degrees" in a graph. The mean they call the "local mean" is precisely analogous to the expected degree of RN. The second mean they define is the "global mean" of the graph, which is the mean degree of the collection of all edge endpoints. Note that a vertex can appear multiple times in this collection, specifically it appears as many times as its degree. We note that the global mean is exactly equal to the expected degree of a vertex sampled by a lesser-known sampling method, random edge or RE (Leskovec and Faloutsos 2006; Pal et al. 2019). An edge is sampled at random from the collection of edges in the graph, and one of its two vertex endpoints is selected with uniform probability. The collection of edge endpoints is exactly analogous to a graph's collection of friends that is the basis of the FP, so the FP suffices to prove that $\mathbb{E}[RE] \geq \mathbb{E}[RV]$ and $\mathbb{E}[RE] > \mathbb{E}[RV]$ in all graphs except a regular graph. Of course, as a practical sampling method, RE is often impossible because edges are typically not tracked as an independent collection. Our research is academic in nature, so we analyze results and ignore the practicality of the methods' implementations. Still, it is worth noting that RE is not impossible. Obviously, any online graph has the option to track edges if it would be advantageous to do so. Also, the probabilistic method suggested in Kumar et al. (2018) is another way of achieving RE, even without an independent collection of edges.

We can express the expected degree of a vertex sampled by RE as

$$\mathbb{E}[RE] = \frac{1}{m} \sum_{e(u,v)\in E} \frac{d_u + d_v}{2} \tag{3}$$

where m is the number of edges in the graph.

RN Versus RE

Kumar et al. (2018) prove that either of their two means can be greater than the other, so by direct extension, both $\mathbb{E}[RN] > \mathbb{E}[RE]$ and $\mathbb{E}[RE] > \mathbb{E}[RN]$ are possible in different graphs.

A specific focus of our research is the ratios between the different sampling methods, so we establish the equations of the two ratios that relate the exclusive methods.

$$\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = \frac{2m}{n} \frac{\sum_{e(u,v)\in E} \left(\frac{d_u}{d_v} + \frac{d_v}{d_u}\right)}{\sum_{e(u,v)\in E} (d_u + d_v)} \tag{4}$$

And the inverse

$$\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]} = \frac{n}{2m} \frac{\sum_{e(u,v)\in E} (d_u + d_v)}{\sum_{e(u,v)\in E} \left(\frac{d_u}{d_v} + \frac{d_v}{d_u}\right)}$$

Theorem 1 $\frac{\mathbb{E}[\boldsymbol{RN}]}{\mathbb{E}[\boldsymbol{RE}]} \leq \frac{2\boldsymbol{m}}{\boldsymbol{n}}$.

Proof Every edge contributes a value in the form of $\frac{\boldsymbol{a}}{\boldsymbol{b}} + \frac{\boldsymbol{b}}{\boldsymbol{a}}$ to the numerator of the second term in Eq. 4, and a value in the form of $\boldsymbol{a} + \boldsymbol{b}$ to the denominator.

$$\frac{a}{b} + \frac{b}{a} = \frac{\mathrm{a}^2 + \mathrm{b}^2}{\mathrm{ab}} \leq \frac{a^2 b + b^2 a}{ab} = a + b$$

□

Corollary 1 $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} < \frac{2m}{n}$ *in all graphs with a single vertex v with* $d_v > 1$.

Proof There exists at least one edge (u, v) with $d_u > 1$. If $a > 1$ and $b \geq 1$ then.

$$a^2 + b^2 < a^2 b + b^2 a$$

□

Inclusive random sampling

We are proposing a tweak to both RN and RE where an informed decision is made that assures the higher-degree vertex of the two vertices being considered is the one that is selected.

Inclusive RN (IRN)

Recall that in RN we sample a vertex at random, then sample a neighbor from among its neighbors and select it instead. In IRN, we learn the degree of both the initially

sampled vertex and the sampled neighbor, and we retain the vertex of higher degree. This is essentially a correction for the outlying cases where the initial vertex has a higher degree than the selected neighbor, in other words the individual samplings where RV would have been superior to RN.

To calculate the expected degree, we can rewrite Eq. 1 as

$$\mathbb{E}[IRN] = \frac{1}{n}\sum_{v\in V}\sum_{u\in \mathrm{N}_v}\frac{\max(d_u,d_v)}{d_v}$$

We can also rewrite Eq. 2 as

$$\mathbb{E}[IRN] = \frac{1}{n}\sum_{e(u,v)\in E}\frac{\max(d_u,d_v)}{d_v}+\frac{\max(d_u,d_v)}{d_u} \tag{5}$$

To make the notation simpler, we stipulate that an edge expressed as $e(u,v)$ always places the endpoint vertices in descending order of degree, in other words $d_u \geq d_v$. This allows us to rewrite Eq. 5 more simply as

$$\mathbb{E}[IRN] = \frac{1}{n}\sum_{e(u,v)\in E}\left(\frac{d_u}{d_v}+1\right) \tag{6}$$

IRN versus RN

Clearly $\mathbb{E}[IRN] \geq \mathbb{E}[RN]$ and the two values are only equal in a perfectly assortative graph. Equations 6 and 2 can be used to establish the difference between IRN and RN as $\mathbb{E}[\boldsymbol{IRN}] \leq \mathbb{E}[\boldsymbol{RN}] + \frac{\boldsymbol{m}(\boldsymbol{n}-2)}{\boldsymbol{n}(\boldsymbol{n}-1)}$.

We next examine the ratio between the two.

Theorem 2 $\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]} \leq \frac{\sqrt{2}+1}{2}$.

Proof Using Eqs. 6 and 2 we can express the ratio as

$$\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]} = \frac{\sum_{e(u,v)\in E}\left(\frac{d_u}{d_v}+1\right)}{\sum_{e(u,v)\in E}\left(\frac{d_u}{d_v}+\frac{d_v}{d_u}\right)}$$

We seek to maximize an expression in the form of

$$\frac{\frac{x}{y}+1}{\frac{x}{y}+\frac{y}{x}}, x \geq y$$

Differentiating the function gives

$$\frac{d}{dx} = \frac{\left(x^2+y^2\right)(2x+y)-2x(x^2+xy)}{\left(x^2+y^2\right)^2}$$

And setting this expression to 0 gives two extremal points at $x = y(1 \pm \sqrt{2})$. Because $x \geq y$, we only consider $x = y\left(1 + \sqrt{2}\right)$, and the sign of the second derivative at this point confirms that this is a maximal value. We can therefore maximize the ratio as

$$\max\left(\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]}\right) = \frac{\sqrt{2}+1+1}{\sqrt{2}+1+\frac{1}{\sqrt{2}+1}} = \frac{\sqrt{2}+1}{2}$$

□

Theorem 2 is a tight upper bound. Consider a complete bipartite graph with k vertices on one side and $\sim k(\sqrt{2}+1)$ vertices on the other. The ratio approximates

$$\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]} \cong \frac{\sum_{e(u,v)\in E} \frac{k\left(\sqrt{2}+1\right)}{k} + 1}{\sum_{e(u,v)\in E} \frac{k\left(\sqrt{2}+1\right)}{k} + \frac{k}{k\left(\sqrt{2}+1\right)}} = \frac{2+\sqrt{2}}{\sqrt{2}+1+\frac{1}{\sqrt{2}+1}} = \frac{\sqrt{2}+1}{2}$$

Inclusive RE (IRE)

Recall that RE involves selecting an edge at random from the edges of a graph and then selecting one of the two endpoints at random. In IRE, we learn the degree of both endpoints and select the one of higher degree. In RN, inclusive sampling is a correction for outlying cases, blindly selecting the neighbor does give a higher expected degree. In RE, on the other hand, selecting the lower-degree vertex is not an outlying case, it occurs with equal probability. The correction of inclusive sampling, therefore, is intuitively stronger.

We can rewrite Eq. 3 as

$$\mathbb{E}[IRE] = \frac{1}{m}\sum_{e(u,v)\in E} d_u \tag{7}$$

IRE versus RE

As with RN, it is obvious that inclusivity only increases the expected degree, $\mathbb{E}[IRE] \geq \mathbb{E}[RE]$, and the values are only equal in a perfectly assortative graph. We again consider the improvement both in terms of the maximum difference between the two expected degrees and the maximum ratio between the two. Using Eqs. 7 and 3, it is not difficult to establish the difference as:

$$\mathbb{E}[\boldsymbol{IRE}] \leq \mathbb{E}[\boldsymbol{RE}] + \frac{\boldsymbol{n}}{2} - 1$$

It is interesting to note that the star graph of n vertices maximizes the difference over all graphs of n vertices because every edge achieves the maximum amount.

We next establish the ratio between IRE and RE as follows:

Theorem 3 $\frac{\mathbb{E}[\boldsymbol{IRE}]}{\mathbb{E}[\boldsymbol{RE}]} < 2$.

Proof Using Eqs. 6 and 3

$$\frac{\mathbb{E}[IRE]}{\mathbb{E}[RE]} = \frac{\frac{1}{m}\sum_{e(u,v)\in E} d_u}{\frac{1}{m}\sum_{e(u,v)\in E} \frac{d_u+d_v}{2}}$$

The ratio for any edge is

$$\frac{d_u}{\frac{d_u+d_v}{2}} = \frac{2d_u}{d_u+d_v}$$

And clearly $2d_u < 2(d_u + d_v)$.

□

Here the star graph demonstrates that the bound is tight because it minimizes d_v for every edge, and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[RE]}$ approaches the maximum possible value of 2 as n increases.

It is interesting to note that the $\frac{\mathbb{E}[IRN]}{\mathbb{E}[RN]}$ ratio for the star graph approaches 1 as n increases. This stark contrast again draws attention to the difference in the natures of the corrections achieved by IRN and IRE. As noted, IRN corrects for an outlying case, in the star graph the case of initially selecting the center which occurs with probability $\frac{1}{n}$. However, IRE corrects more broadly for the case of selecting the lower-degree endpoint of any edge, which in the star graph translates to a .5 probability of selecting a leaf vertex.

IRN versus IRE

We now perform a direct comparison between the two inclusive methods themselves. We first establish that either ratio can grow without bound and then consider possible bounds on the number of vertices required to achieve a desired ratio. It is important to note that Theorems 2 and 3 establish that the improvement of inclusive sampling over exclusive sampling in both IRN and IRE is bound by a constant factor. Therefore, in order to prove that either ratio can grow without bound, it suffices to prove that the exclusive ratios $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ and $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ can both grow without bound.

In order to do this, we construct pathological graphs that accentuate the strengths of each method vis-à-vis the other.

The $\frac{\mathbb{E}[\boldsymbol{RN}]}{\mathbb{E}[\boldsymbol{RE}]}$ and $\frac{\mathbb{E}[\boldsymbol{IRN}]}{\mathbb{E}[\boldsymbol{IRE}]}$ ratios are unbounded

In order to strengthen RN vis-à-vis RE, we construct a graph comprised of two separate subgraphs. One subgraph is a clique of c vertices and the second is a star of s vertices, see Fig. 2. We select values for c and s so that the star has more vertices than the clique, but the clique has more edges than the star. The degree of the center of the star is highest degree of the graph, and RN is more likely to select this vertex because the majority of vertices in the graph are the leaves of the star that connect to this center vertex. RE,

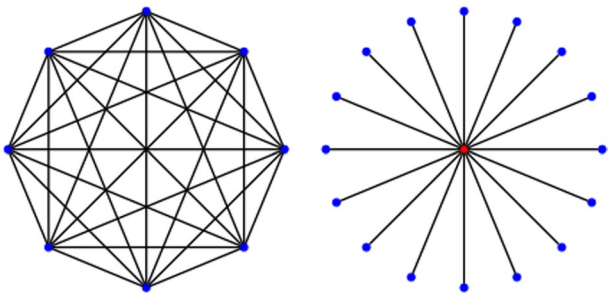

Fig. 2 A graph where RN outperforms RE

on the other hand, is more likely to select one of the vertices in the clique, which are of lower degree than the center of the star, because the majority of edges are in the clique.

In this construction, the ratio $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ is unbounded. We can calculate $\mathbb{E}[RN]$ as

$$\mathbb{E}[RN] = \frac{c(c-1) + (s-1)^2 + 1}{c+s} \tag{8}$$

And $\mathbb{E}[RE]$ as

$$\mathbb{E}[RE] = \frac{c(c-1)^2 + s(s-1)}{c(c-1) + 2(s-1)} \tag{9}$$

Therefore, the ratio is

$$\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = \left(\frac{c(c-1) + (s-1)^2 + 1}{c+s}\right)\left(\frac{c(c-1) + 2(s-1)}{c(c-1)^2 + s(s-1)}\right)$$

Set $c = x^2$ and $s = x^3$. As x increases, the expression approaches

$$\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = \left(\frac{x^6}{x^3}\right)\left(\frac{x^4}{x^6}\right)$$

And this expression can clearly be made arbitrarily large by increasing x.

Bounding $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ as a function of n

Having established that the ratio is unbounded, an interesting question to explore is how many vertices would be required to achieve a desired value. As one possibility, we offer a simple bound for this construction of $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = \boldsymbol{\Omega}\left(\boldsymbol{n}^{\frac{1}{3}}\right)$.

We have set $c = x^2$ and $s = x^3$ which means $n = x^3 + x^2$. If Eq. 8 is rewritten in terms of x, it is easy to prove that $\mathbb{E}[RN] > (x^2+1)(x-1)$. If Eq. 9 is rewritten in terms of x, it is easy to prove that $\mathbb{E}[RE] < 2(x^2-1)$. We can therefore say that

$$\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} > \frac{\left(x^2+1\right)(x-1)}{2\left(x^2-1\right)} > \frac{x+1}{2} - 1$$

Because $n = c + s = x^3 + x^2$, $x + 1 > n^{\frac{1}{3}}$, so we can conclude $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = \Omega\left(n^{\frac{1}{3}}\right)$.

As we have noted, because $\mathbb{E}[IRN] \geq \mathbb{E}[RN]$ and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[RE]} < 2$, the results apply to $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$ as well, that is $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$ can grow without bound and has a possible lower bound of $\Omega\left(n^{\frac{1}{3}}\right)$.

The $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ ratios are unbounded

We now take the opposite approach and provide a construction that strengthens RE vis-à-vis RN. The first subgraph is again a clique of size c. The second subgraph is a set of s degree-1 vertices joined by $\frac{s}{2}$ edges. We once again put the majority of edges in the clique, and the majority of vertices in the set of edges, see Fig. 3.

Once again, RE is more likely to select a vertex from the clique while RN is more likely to select a vertex from the collection of edges. However, in this construction, the vertices in the clique are the max-degree vertices in the graph, while the vertices in the other subgraph are all degree-1 so $\mathbb{E}[RE] > \mathbb{E}[RN]$.

In this construction the ratio $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ is unbounded. We can calculate $E[RE]$ as follows

$$\mathbb{E}[RE] = \frac{c(c-1)^2 + s}{c(c-1) + s}$$

And the value of $\mathbb{E}[RN]$ is

$$\mathbb{E}[RN] = \frac{c(c-1) + s}{c + s}$$

And therefore

$$\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]} = \frac{\left(c(c-1)^2 + s\right)(c+s)}{(c(c-1)+s)^2} \tag{10}$$

This expression expands to

$$\frac{s^2 + \left(c^3 - 2c^2 + 2c\right)s + c^4 - 2c^3 + c^2}{s^2 + \left(2c^2 - 2c\right)s + c^4 - 2c^3 + c^2}$$

For any fixed s, increasing c increases the ratio, so values of s and c can be selected to achieve any ratio.

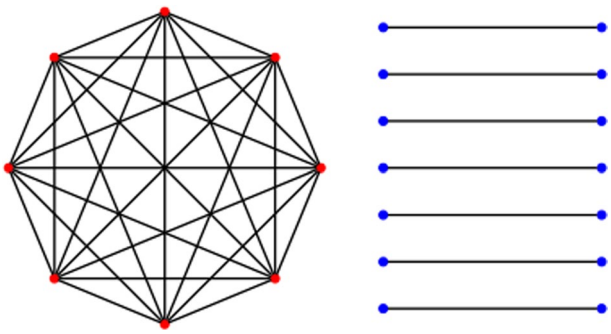

Fig. 3 A graph that favors RE over RN

Bounding $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ as a function of n

Here we can propose a simple lower bound on n as follows. Set $s = c(c-1)$, so $n = c + c(c-1) = c^2$. Rewriting Eq. 10 in terms of c gives

$$\frac{\big(c(c-1)^2 + c(c-1)\big)(c + c(c-1))}{(c(c-1) + c(c-1))^2} = \frac{c^2}{4(c-1)} = \Omega\Big(n^{\frac{1}{2}}\Big)$$

In this construction, extending the results to inclusive sampling is even easier because the graph is perfectly assortative. Therefore $\mathbb{E}[IRE] = \mathbb{E}[RE]$ and $\mathbb{E}[IRN] = \mathbb{E}[RN]$ so $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ is also unbounded and has a possible lower bound of $\Omega\Big(n^{\frac{1}{2}}\Big)$.

$\frac{\mathbb{E}[IRN]}{\mathbb{E}[RE]}$ *and* $\frac{\mathbb{E}[IRE]}{\mathbb{E}[RN]}$

We note two obvious corollaries regarding the ratios between the inclusive methods as bounded by their exclusive counterparts. The corollaries are derived from Theorems 2 and 3.

Corollary 2 $\frac{\mathbb{E}[RN]}{2\mathbb{E}[RE]} < \frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]} \leq \frac{\left(\sqrt{2}+1\right)\mathbb{E}[RN]}{\mathbb{E}[RE]}$

Corollary 3 $\frac{\mathbb{E}[RE]}{(\sqrt{2}+1)\mathbb{E}[RN]} \leq \frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]} < \frac{2\mathbb{E}[RE]}{\mathbb{E}[RN]}$

Random sampling in trees

Trees present an interesting challenge for analyzing these sampling methods. The ratio $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ is not unbounded in trees, a strict bound of 2 is easily proven. If the goal is to maximize $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$, recall that the pathological examples of the previous section included subgraphs that were cliques in order to increase the likelihood of RE selecting one of the vertices of the subgraph. In trees of course, it is impossible to saturate any part of the graph with edges.

$\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ and $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$

We first establish a simple bound on the $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ ratio in trees. Replacing m with $n-1$ in Corollary 1 gives:

Corollary 4 *In all trees,* $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} < \frac{2(n-1)}{n}$, *so* $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} < 2$.

Note that the bound is strict, because it is only possible to use Theorem 1 in a tree of two vertices where $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]} = 1$.

It is interesting to note that $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ in the star graph has the same upper bound, so again the bound is tight and it suggests that the star graph of size n maximizes the ratio $\frac{\mathbb{E}[RN]}{\mathbb{E}[RE]}$ over all trees of size n.

We can easily prove the same bound for $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$. We can express $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$ in trees as

$$\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]} = \frac{n-1}{n}\frac{\sum_{e(u,v)\in E}\frac{d_u}{d_v} + 1}{\sum_{e(u,v)\in E} d_u}$$

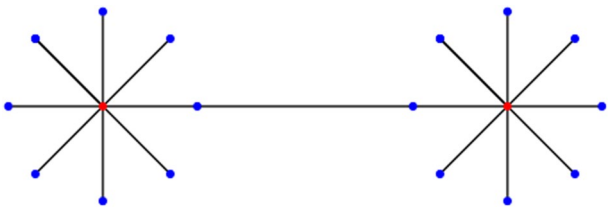

Fig. 4 A graph tree where $\mathbb{E}[IRN] > \mathbb{E}[IRE]$

For any edge $e(u, v)$, the term $\frac{\frac{d_u}{d_v}+1}{d_u} \leq 2$, so the numerator cannot be more than twice the denominator and the inequality is strict because of the first term $\frac{n-1}{n}$.

However here, the star graph fails to achieve the value of the bound because in the star graph $\mathbb{E}[IRN] = \mathbb{E}[IRE]$. In fact, it is not simple to prove the possibility of $\mathbb{E}[IRN] > \mathbb{E}[IRE]$ in trees because of the aforementioned inability to strengthen RE with additional edges. But it is possible as we demonstrate with the example in Fig. 4.

Start with two stars of size c and add a single edge connecting one leaf from each.

$$\begin{aligned}
\mathbb{E}[IRN] &= \frac{2c^2 + c + 2}{2c + 2} \\
\mathbb{E}[IRE] &= \frac{2c^2 + 2}{2c + 1} \\
\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]} &= \left(\frac{2c^2 + c + 2}{2c + 2}\right)\left(\frac{2c + 1}{2c^2 + 2}\right) \\
&= \frac{4c^3 + 4c^2 + 5c + 2}{4c^3 + 4c^2 + 4c + 4}
\end{aligned}$$

$\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ are unbounded in trees

While the $\frac{\mathbb{E}[IRN]}{\mathbb{E}[IRE]}$ ratio is bounded in trees, $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ is still unbounded. We present a construction here that proves $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ and $\frac{\mathbb{E}[IRE]}{\mathbb{E}[IRN]}$ are unbounded even in trees.

Attach c children to a root vertex. For each of the c children, attach $s - 1$ children that are leaves, so that the degrees of the internal vertices are s, see Fig. 5.

$$\mathbb{E}[RE] = \frac{c + s^2 + s - 1}{2s}$$

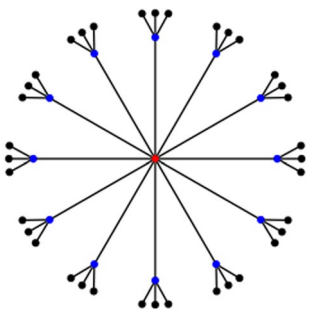

Fig. 5 A construction where $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ is unbounded

$$\mathbb{E}[RN] = \frac{\frac{1}{s}c^2 + \left(s^2 - s + 1 - \frac{1}{s}c\right) + s}{sc + 1}$$

For a fixed s, set $c \gg s$. $\mathbb{E}[$ RE$]$ approaches $\frac{c}{2s}$ and $\mathbb{E}[$ RN$]$ approaches $\frac{c}{s^2}$. So we can say

$$\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]} \approx \left(\frac{c}{2s}\right)\left(\frac{s^2}{c}\right) = \frac{s}{2}$$

Which grows without bound as s increases.

Bounding $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]}$ as a function of n

We again offer a simple possible bound based on our construction. An obvious lower bound on $\mathbb{E}[RE]$ is $\mathbb{E}[RE] > \frac{c}{2s}$. We can express an upper bound of $\mathbb{E}[RN] < \frac{c}{s^2} + s$ if we assume $c > 1$ and subtract 1 from the denominator. If we assume $s^3 < c$, then $\mathbb{E}[RN] < \frac{2c}{s^2}$ and therefore

$$\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]} > \left(\frac{c}{2s}\right)\left(\frac{s^2}{2c}\right) = \frac{s}{4}$$

The number of vertices is $n = cs + 1$, and we are assuming $s^3 < c$, so we can approximate a bound of $\frac{\mathbb{E}[RE]}{\mathbb{E}[RN]} = \Omega\left(n^{\frac{1}{4}}\right)$.

Experimental analysis

We now present some results of experimentation in synthetic graphs and the graphs of real-world networks. For synthetic graphs we use the well-known Erdős Réyni (ER) (Erdős and Rényi 1959) and Barabási Abert (BA) (Barabási and Albert 1999) models, and we examined the graphs of real-world networks from the Koblenz Network Collection (Kunegis 2013).

Synthetic graphs

In both ER and BA graphs an interesting trend emerges. In both types, as would be expected, $\mathbb{E}[RN] > \mathbb{E}[RV]$ and $\mathbb{E}[RE] > \mathbb{E}[RV]$ as the graphs will almost certainly contain an edge between two vertices of different degree. The gains for both methods over RV are modest in ER graphs but significant in BA graphs. In ER graphs, RN is always minimally better than RE. In BA graphs this is almost always true as well, but when the edge count is very high RE outperforms RN. This is seemingly consistent with our analysis of the pathological example in Fig. 2. The increase in edge count likely increases substructures that resembles cliques instead of stars and this boosts the performance of RE. RN's strong performance in BA graphs is likely linked to the traits of the power-law distribution and assortativity. As we discuss in subsequent sections, the power-law distribution typically causes some amount of disassortativity, and this in turn strengthens RN.

Inclusive sampling in synthetic graphs

The inclusive sampling reveals an interesting result which is consistent with the theoretical bounds we have established. Unsurprisingly, the assumptions $\mathbb{E}[IRN] > \mathbb{E}[RN]$ and

Table 1 Sampling method results for ER/BA Graphs, n = 6000

RV	Erdős Rényi Graphs, n = 6000				Barabási Albert Graphs, n = 6000			
	RN	RE	IRN	IRE	RN	RE	IRN	IRE
6	6.9952	6.9946	7.9227	8.361	19.54	17.68	21.34	29.63
10	10.9883	10.9882	12.3023	12.755	27.87	26.18	30.7	42.76
16	16.973	16.9714	18.7509	19.2119	38.9	37.43	43.24	59.46
30	30.922	30.9212	33.525	33.9967	63.89	62.75	71.64	96.63
60	60.6866	60.6864	64.5381	65.0121	113.3	112.55	128.18	167.78
129	129.5657	129.565	135.4022	135.8766	216.32	216.42	246.99	310.69

$\mathbb{E}[IRE] > \mathbb{E}[RE]$ hold. While it is almost always true that $\mathbb{E}[RN] > \mathbb{E}[RE]$, it is always true that $\mathbb{E}[IRE] > \mathbb{E}[IRN]$. This again seems to reflect on the more corrective nature of IRE, and it also follows naturally from the greater potential indicated by the bound of 2 in Theorem 3 versus the smaller bound of ~ 1.21 of Theorem 2. The results are summarized in Table 1 below.

Real-world networks

We examined 1072 networks from the Koblenz Network Collection (Kunegis 2013) to see the effects of the four sampling methods. We find that $\mathbb{E}[RN] > \mathbb{E}[RE]$ in 93% of the networks, yet $\mathbb{E}[IRE] > \mathbb{E}[IRN]$ in 43%. The average gain of IRN versus RN is

Table 2 Method comparisons in real-world networks by category

Category	Pct RN > RE (%)	Pct IRN > IRE (%)	IRN/RN	IRE/RE
Affiliation	100	17	1.05	1.68
Animal	75	0	1.09	1.13
Authorship	99	67	1.01	1.94
Citation	50	0	1.08	1.58
Cocitation	0	0	1.1	1.47
Communication	83	25	1.04	1.7
Computer	64	0	1.07	1.60
Feature	83	50	1.02	1.88
Human Contact	86	14	1.12	1.31
Human Social	55	0	1.12	1.21
Hyperlink	71	14	1.02	1.84
Infrastructure	48	0	1.1	1.2
Interaction	81	62	1.04	1.71
Lexical	67	33	1.08	1.66
Metabolic	75	0	1.07	1.59
Misc	67	0	1.08	1.55
Neural	100	0	1.11	1.45
Online Contact	75	13	1.03	1.69
Rating	100	57	1.02	1.87
Social	71	31	1.03	1.76
Software	100	67	1.003	1.98
Text	83	0	1.04	1.58
Trophic	100	0	1.14	1.33

102.3%, while the average gain of IRE versus RE is a staggering 186%. This is especially significant in light of the bound of 2 in Theorem 3.

We also calculate these results for the different network categories of the collection. The results are summarized in Table 2. $\mathbb{E}[RN] > \mathbb{E}[RE]$ in the majority of networks in all but three categories, and the mean percent over all categories where this is true is 72.8%. $\mathbb{E}[IRE] > \mathbb{E}[IRN]$ in a majority of networks in all but three categories (note that these are not the same three categories where $\mathbb{E}[RE] > \mathbb{E}[RN]$), and the mean percent over all categories where this is true is 82.2%. The modest gains of IRN over RN are roughly consistent over all categories, while the gain of IRE over RE ranges from 1.13 to 1.98.

The influence of degree-homophily and the power-law

In Novick and Bar-Noy (2021, 2022) we outlined an analysis of how the power-law distribution that defines BA graphs and is a common trait of many real-world graphs (Barabási and Albert 1999) typically implies an amount of disassortativity, and this in turn strengthens RN. The relatively low count of high-degree vertices cannot satisfy their total edge endpoints without connecting to some of the low-degree vertices, and this disassortativity strengthens RN because the vertex initially sampled, which is likely of low-degree, has some significant likelihood of being connected to a high-degree vertex that may be selected by RN. This is a significant difference between ER and BA graphs. Both are known to be non-assortative (Newman 2002), but research has shown that in ER graphs this non-assortative nature is more homogeneous, while in BA graphs it results from an aggregate measure of two sharply contrasting types of connections, some assortative and some disassortative (Bertotti and Modanese 1806).

This phenomenon was explored by Kumar et al. (2018) as well. The authors introduced a new measure, 'inversity', and showed how its sign perfectly predicts which of RN and RE would have the higher expected degree. While this is not true of assortativity, the correlation between inversity and assortativity is very high, and our purpose is only to demonstrate the effect of degree-homophily in general, so we based our results on assortativity. Here we extend those results and examine their application on inclusive sampling.

Power-law distribution

Our first experiment checks the effect of the power-law on all sampling methods. Recall the equation used in the Barabási Albert algorithm (Barabási and Albert 1999) for determining the vertices to which a new vertex connects

$$p(v_i) = \frac{d_{v_i}}{\sum_{v \in V} d_v}$$

This motivates the preferential attachment that causes the power-law distribution, the probability of a vertex being selected is directly proportional to its degree.

It is possible to generalize the equation with a parameter α as follows

$$p(v_i) = \frac{d_{v_i}^{\alpha}}{\sum_{v \in V} d_v^{\alpha}}$$

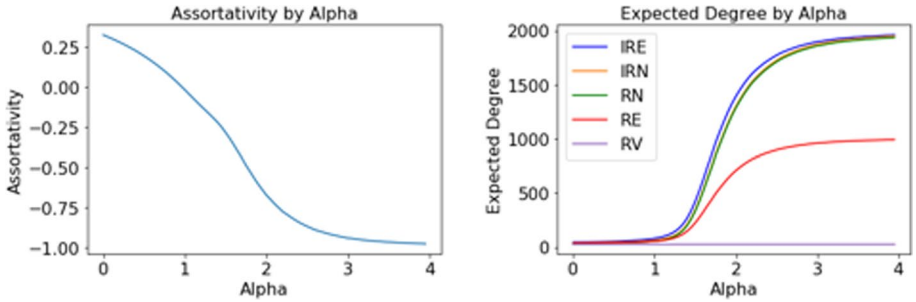

Fig. 6 Assortativity and sampling expectations for tweaked BA graphs

The original equation has $\alpha = 1$. It is possible to weaken the preferential attachments by setting $\alpha < 1$ and to strengthen it by setting $\alpha > 1$.

We generated BA graphs with varying values of α and tracked the results on the sampling methods. As demonstrated in Fig. 6, the increase in α decreases degree-homophily as measured by assortativity. This decrease increases the values of all four sampling methods. It interesting to note that RE outperforms RN for smaller values of α, but as α reaches the original value of 1 and surpasses it, RN becomes the superior method. However, we again see the phenomenon that inclusive sampling corrects RE so much more than RN and IRE is the stronger method of the two inclusive sampling methods.

Rewiring for assortativity

Our final experiment examines the effects of assortativity more directly. Using the technique presented in Mieghem et al. (2010), Xulvi-Brunet and Sokolov (2004) among others, we take ER and BA graphs, and rewire them to both decrease and increase assortativity, tracking the expected degree of the four sampling methods. The results are shown in Fig. 7.

It is important to note that rewiring preserves the degree sequence of a graph even while it changes characteristics such as degree-homophily. This is a contrast to the previous experiment where tweaking the power-law distribution actually changes the degree sequence.

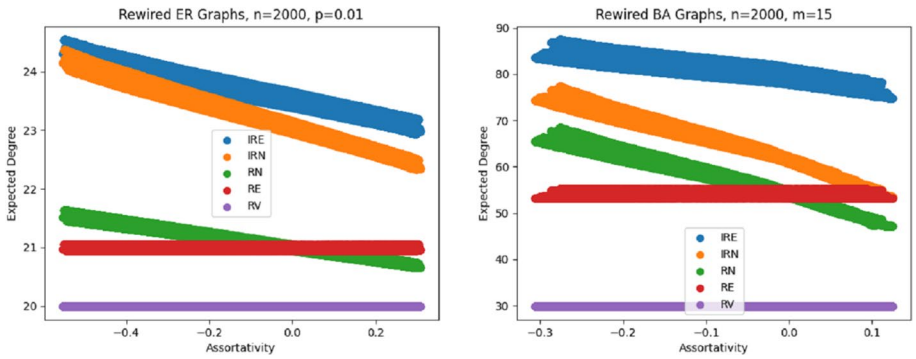

Fig. 7 Sampling expectations for rewired ER and BA graphs

RE is purely a function of the degree sequence and, as such, the results do not change. RN, on the other hand, increases markedly with disassortativity. It is also interesting to note that the two intersect near the value of perfect non-assortativity. Although assortativity is not as precise as inversity, this result is still in line with the results of Kumar et al., as 0 inversity and 0 assortativity will be very close due to the strong correlation between the two values.

The results on inclusive sampling are telling. Firstly, the superiority of the inclusive methods is evident. Secondly, we see again that IRE is superior to IRN. And lastly, we see that although increasing assortativity diminishes the strengths of both inclusive methods, it seems to weaken IRN more significantly than IRE, another point in favor of IRE as a sampling method.

Conclusion and future research directions

This paper has introduced the idea of inclusive random sampling and applied it to the well-known random neighbor sampling method as well as the less-known random edge sampling method. We studied both the original, exclusive versions of these methods along with the new, inclusive ones. We have proven that either version's ratio to the other can grow without bound and provided additional interesting bounds on the methods' performances vis-à-vis each other and their exclusive counterparts. We also conducted a study in the specific case of trees, noting which general results apply equally to trees and which do not.

Through experimentation on synthetic and real-world graphs, we established the usefulness of inclusive sampling as a practical method. We have many findings to reflect on this practical application of our research, most prominent among them the fact that IRE is often superior to IRN, even when RN is superior to RE. This suggests a potential value in tracking edges of a graph when high-degree random sampling is important.

We have also shown the relationship between preferential attachment and degree-homophily on one hand and inclusive sampling on the other. These findings can aid in the analysis of a particular graph to determine which sampling method is likely to yield the highest expectation of degree. Of course, there are other graph traits and phenomena that may be linked to the performance of these sampling methods. We believe there is a lot of potential to explore what other graph types and structures could influence these outcomes. In addition, there could be other factors that influence the decision, such as the cost of tracking edges, that could be taken into account. We hope to explore these concepts further and continue to contribute to the understanding of how these sampling techniques work and how best to utilize them.

Abbreviations

RN	Random neighbor sampling
RE	Random edge sampling
IRN	Inclusive random neighbor sampling
IRE	Inclusive random edge sampling
FP	Friendship paradox
ER	Erdős Rényi (random graph)
BA	Barabási Albert (random graph)

Author contributions
A.B.N. suggested many of the research avenues, provided the pathological graph constructions and the bounds they prove, and proved the upper bound on E[IRN]/E[RE]. Y.N. established the equations, proved other bounds, provided the demonstration of E[IRN] > E[IRE] in trees, and conducted the experimental analyses. Y.N. wrote all text and prepared all figures. Both authors reviewed and approve of the manuscript.

Funding
Not applicable.

Availability of data and materials
Sample real-world networks for some experiments were taken from the Koblenz network collection, http://konect.uni-koblenz.de/.

Declarations

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

Ethical approval
Not applicable.

Received: 30 October 2022 Accepted: 3 August 2023
Published online: 04 September 2023

References

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
Bertotti ML, Modanese G (2018) The bass diffusion model on finite barabasi-albert networks. Phys Soc. arXiv:1806.05959
Cohen R, Havlin S, Ben-Avraham D (2003) Efficient immunization strategies for computer networks and populations. Phys Rev Lett 91:24
Christakis NA, Fowler JH (2010) Social network sensors for early detection of contagious outbreaks. PLoS ONE 5(9):e12948
Erdős P, Rényi A (1959) On random graphs I. Publicationes Mathematicae 6:290
Feld S (1991) Why your friends have more friends than you do. Am J Soc 96(6):1464–1477
Kumar V, Krackhardt D, Feld S (2018) Network interventions based on inversity: leveraging the friendship paradox in unknown network structures. https://vineetkumars.github.io/Papers/NetworkInversity.pdf
Kunegis J (2013) KONECT, The Koblenz network collection. http://konect.cc/
Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: 12th ACM SIGKDD international conference on knowledge discover and data mining (2006)
Malliaros FD, Rossi MEG, Vazirgiannis M (2016) Locating influential nodes in complex networks. Sci Rep 6(1):19307
Momeni N, Rabbat MG (2018) Effectiveness of alter sampling in social networks. https://arxiv.org/abs/1812.03096v2 (2018)
Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701
Novick Y, Bar-Noy A (2020) Finding high-degree vertices with inclusive random sampling. In: International conference on complex networks and their applications. Springer, Cham
Novick Y, Bar-Noy A (2021) A fair-cost analysis of the random neighbor sampling method. In: International conference on complex networks and their applications. Springer, Cham (2021)
Novick Y, Bar-Noy A (2022) Cost-based analyses of random neighbor and derived sampling methods. Appl Netw Sci 7(1):34
Pal S, Yu F, Novick Y, Swamin A, Bar-Noy A (2019) A study on the friendship paradox—quantitative analysis and relationship with assortative mixing. Appl Netw Sci 4:71
Strogatz S (2012) Friends you can count on, NY Times 9/17/2012. https://opinionator.blogs.nytimes.com/2012/09/17/friends-you-can-count-on/
Van Mieghem P, Wang H, Ge X, Tang S, Kuipers FA (2010) Influence of assortativity and degree-preserving rewiring on the spectra of networks. Eur Phys J B 76:643–652
Xulvi-Brunet R, Sokolov IM (2004) Reshuffling scale-free networks: from random to assortative. Phys Rev 70:066102

Publisher's Note