Performance of random walks and sampling for graph search

A Thesis

Submitted to the Faculty

of

Drexel University

by

Jonathan Stokes

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

September 2018

Dedications

To all those who have and continue to put up with me.

Acknowledgments

I would like to thank a number of people who have supported my work. I would not be writing this without having known through out my PhD. that in the worst case I had at least one couch to sleep on. For this I would like to thank my immediate family Cherie Aiello and Godi Fischer, Russell Stokes and Mary Finnegan, Laurel Stokes, and Alea Stokes.

I would also like to thank my first advisor Prof. Mark Hempstead who gave me the opportunity to work in his research lab, many of the skills I learned in his lab I continue to use on a daily basis. The students in Prof. Hempstead's also guided me through my first years at Drexel including Steven Battle, Siddharth Nilakantan, Rizwana Begum, Jason Palaszewski, and Tianyun Zhang. I would also like to thank Prof. John Walsh for introducing me to stochastic processes as well as Solmaz Torabi and Owen Mayer for taking many of his courses with me.

Of course I would not be writing this without the guidance of my advisor Prof. Steven Weber. I did not appreciate what I was getting myself into when I joined his lab and although not all our research projects have gone smoothly, I have learned and accomplished more under his guidance then I thought myself capable of. In addition, I must thank Prof. Harish Sethu who co-sponsored my research with Prof. Weber. I would also like to thank Prof. Weber's students Jeffery Wildman, Bradford Boyle, and Nan Xie for always giving me sound advice and in particular Ni An who has put up with me for the entirety of my PhD.

Finally I would like to thank my committee members Prof. Matthew Stamm, Prof. Jake Williams, Prof. Jaudelice de Oliveira, and Prof. Nagarajan Kandasamy for taking the time to review my work, if there is one thing I have learned in the course of this PhD. it is the value of time.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation and Contributions

Graphs represent connections among a group of entities, whether those connections and entities are the communication links between hubs in a computer network, related values between records in a database, or friendships among people in a social network. With recent advances in communication and computation technology, and the subsequent ability to easily collect and transmit data, it is not uncommon to find in practice extremely large graphs, in both the number of entities (say, billions or more) and the number of connections. Even more, it is often the case, particularly with social networks, that these graphs are neither static (in time) nor local (in space), meaning it is not possible, or at least not practical, to store the graph on desktop computer.

With this growth in sizes and scale, it is important to understand the computational effort incurred in seeking answers to questions of interest about large graphs. These questions may take a wide variety of forms — one may be interested in studying how disease or information spreads through a population, how to quickly return a database query about the number of records with a given property, or how to efficiently find the social media user with the most-liked cat photograph. In any case, answering these questions requires analyzing the graph. Much (but not all) of this thesis focuses on the particular question of identifying a vertex with maximum, or near-maximum, degree. A natural and nontrivial motivation for this question is that high degree vertices govern the robustness of a network and how quickly information (or a contagion) spreads through it.

In contexts like the one mentioned above, where the graph is not available locally, the analysis of the graph requires that the graph's connections be investigated by submitting a sequence of (randomized) queries. We say that a graph query consists of a vertex label, and the query response is to provide the value of some property of interest of the vertex, the neighbors of that vertex, and (possibly) the property of interest for each of the neighbors. Queries are assumed to be submitted sequentially until the question at hand has been answered, and the cost of answering the question

of interest about the graph is measured either as the number of queries that are submitted, or as the number of vertices for which the property of interest must be computed. Given our assumption that a query returns the values of the property of interest of the neighbors of the query vertex, we are led naturally to the notion of a *star*, where a star consists of a vertex and its one-hop neighbors.

It is of interest to understand how the cost, defined above, to answer these questions about graphs will depend upon both the question being asked, as well as on various summary properties about the graph, including the size (number of edges) and order (number of vertices). Such an understanding will assist with determination of the computational and memory resources and time required for answering various types of questions on various types of graphs.

We focus our attention on two popular randomized approaches for studying a large graph, namely, *random walk* algorithms and *random sampling* algorithms. Random walk algorithms select the next vertex to be queried among the neighbors of the vertex from the previous query, while graph sampling algorithms select a vertex at random from the graph as a whole. Given that each query returns a star, we are led to the analysis of *star sampling*, in which each sample returns a random vertex and its neighbors.

One important aspect of our investigation in comparing the relative cost of random walk and random sampling is to study the benefit of local information for random walk algorithms. Intuitively, random walk should outperform random sampling if knowledge about the properties of the neighbors of the recent query informs the selection as to which vertex to submit for the subsequent query. In the context of searching for maximum degree vertices, we leverage the notion of *assortativity*, the correlation of the degrees of the endpoints of a randomly selected edge, as a proxy for local information about the "location" of the target vertex. Positively correlated graphs, in which high degree vertices tend to be attached to other high degree vertices, exhibit a "local gradient" which may be prove of use in searching for a maximum degree vertex. The value of this gradient in seeking maximum degree vertices is studied by analyzing the query complexity on graphs using both random walks and random sampling, sweeping over graphs of varying assortativity.

Major questions, and the contributions made in answering them, are listed below.

Questions

1. Of what value is the knowledge of the local properties of a graph in guiding the selection of the next vertex to be queried, as used in random walk graph search?
2. What is the impact of the assortativity on the expected "cost" incurred in finding maximum degree vertices via random walk?
3. What is the expected number of maximum degree vertices in Erdős Rényi (ER) graphs?
4. When using star sampling, in which each graph query returns the properties of the neighborhoods of the vertex, what is the impact on the expected "cost" of sampling with vs. without replacement?
5. When using star sampling, is sampling without replacement always superior in cost to sampling with replacement?
6. Will greedy graph construction algorithms, including wiring a new graph or rewiring an existing graph, succeed in obtaining a graph of maximum assortativity with a specified degree sequence?
7. How accurately may we estimate the expected number of samples required to find a near-maximum value in a dataset, without knowing what the maximum value a priori?

Contributions

1. Analysis of the absorbtion time of a discrete-time Markov chain (DTMC) model of a random walk on a graph suggests an optimal "bias" parameter, which controls the probability that the next step on the walk selects a maximum degree neighbor of the current vertex. We investigate the variation of the optimal bias parameter as a function of the graph assortativity, and show that biased random walks incur lower cost than star sampling on graphs with positive assortativity.
2. The expected number of maximum degree vertices in an ER random graph is arbitrarily close to one in the limit as the graph order grows to infinity, under a natural scaling of the edge probability in the graph order. An implication of this result is that the expected cost of searching for any maximum degree vertex is not significantly less than the expected cost of searching for all maximum degree vertices, at least on large ER graphs.
3. In analyzing star sampling we consider three variants regarding replacement: star sampling with replacement (SS-R), star sampling without replacement of the center vertex (SS-C), and star sampling without replacement of the star (SS-S). We give reasonably accurate expressions for the expected costs of all three variants on an ER graph, and show that, asymptotically in the graph order, all three cost ratios are unity, where cost is the number of required star samples.
4. In analyzing greedy constructions (both wiring or rewiring) of graphs of maximum assortativity with a given degree sequence, we show that several natural greedy rewiring heuristics and an elegant greedy wiring heuristic from the literature may fail to produce such a graph. Even more, the wiring heuristic may fail to produce a graph wtih the target degree sequence.
5. In seeking to predict the maximum value from a long list of independent and identically distributed random values from a sample, we are able to leverage basic results in extreme value theory to identify the sample size needed for the prediction to be sufficiently accurate, when the values are either binomial random variables or Zipf random variables.

1.2 Organization

This thesis is organized as follows.

Chap. 2 introduces a number of central concepts. Specifically, Sec. 2.3 formally introduces degree-biased random walks (BRW) as a graph search mechanism. Given an initial vertex v in a graph $G = (\mathcal{V}, \mathcal{E})$, a BRW iteratively hops to an adjacent vertex u with probability proportional to the degree of vertex u, denoted d_u. This weighted transition probability entails that a BRW is more likely to visit high degree vertices than low degree vertices. Sec. 2.4 introduces the notion of the joint degree matrix $\mathbf{K} = \{K_{j,k}\}$ for $j, k \in [\lambda]^+$ where λ is the maximum degree of graph G. The entries $K_{j,k}$ of $\mathbf{K}$ are the number of edges between vertices of degree j and k in G. Additionally Sec. 2.4 introduces the conditional degree distribution matrix $\mathbf{H} = \{H_{j,k}\}$, the j_{th} row of which is the probability vector $\mathbf{H}_j = \{H_{j,k}\}$ for $k \in [\lambda]^+$, with entries $H_{j,k}$ holding the probability that an edge has a degree k endpoint conditioned on the other endpoint being degree j. Next, graph rewiring algorithms are introduced, as a means to increase or decrease the assortativity α of a graph G Newman [2002]. Having introduced these concepts, Sec. 2.4 gives an approximation for the degree transition matrix $\tilde{\mathbf{P}}_Z$ for the DTMC model of a BRW. The matrix $\tilde{\mathbf{P}}_Z$ is then used to estimate the expected time required for a BRW to reach a maximum degree vertex. In the typical case where the number of distinct degrees found in the graph is significantly smaller than the graph order $n = |\mathcal{V}|$, it follows that the matrix $\tilde{\mathbf{P}}_Z$ is correspondingly smaller than the vertex transition matrix $\mathbf{P}_V$ for a BRW. Therefore, computing the expected time for a BRW to reach a maximum degree vertex from $\tilde{\mathbf{P}}_Z$ (via matrix inversion) is computationally simpler than computing it from $\mathbf{P}_V$. Simulation results are given in Sec. 2.5 for Assortative Erdős Rényi (AER) graphs generated using the rewiring algorithm introduced in Sec. 2.4. These simulations compare the BRW absorption time and the DTMC model of absorption time over graphs whith a range of assortativities. Monte carlo simulations are used to test the ability of the DTMC BRW model to pick the optimal bias coefficient β for a BRW. Finally Sec. 2.5 gives simulation results comparing the absorption time of a BRW, random sampling without replacement (RSW), and star sampling without center replacement (SS-C).

Chap. 3 is divided into two parts. The first, Sec. 3.2, demonstrates that it is difficult to find a maximum degree vertex on account of the fact that, on average, there is a unique such vertex in an Erdős Rényi graph. The second, Sec. 3.3.1, introduces Newman's graph assortativity α, and a rewiring algorithm to construct graphs with a target assortativity α_t. Sec. 3.3.2 uses the graphs constructed to have a target assortativity in simulations which compare the samples required to find a maximum degree vertex $v \in \mathcal{V}_\lambda$ using a BRW, RSW, and star sampling without star replacement (SS-S). The novelty of the second part of Chap. 3 is in the introduction of an estimate of the fraction of strict and non-strict local maxima in a graph G based on its conditional degree distribution $\mathbf{H}$. This section also shows that, in simulations, the average time to absorption of a BRW has a positive, albeit small, correlation with the number of strict local maxima in the graph.

Chap. 4 introduces a more sophisticated self-avoiding BRW-based graph search algorithm entitled self-avoiding walk-jump (SAWJ). SAWJ restarts at local maxima in positively assortative graphs, and at local minima in negatively assortative graphs. This is in contrast with the algorithm introduced by Avrachenkov et al. [2012] which restarts with uniform probability. Sec. 4.4.2 gives an analysis of a simplified version of SAWJ entitled the Walk Jump algorithm (WJ). This analysis only holds for WJ on random AER graphs G_α, as defined in Sec. 4.3.1. The key differences between WJ and SAWJ in Sec. 4.5 is 1) WJ always restarts at local maxima while SAWJ restarts with probability $\beta \in [0, 1]$, and 2) WJ is not self avoiding. However if $\beta = 0$ the time required for WJ to find a maximum degree vertex approximately upper bounds the time SAWJ requires. The analysis of WJ is possible because Sec. 4.3.3 gives an approximation, $\tilde{\mathbf{H}}$, of the conditional degree distribution in a random AER graph G_α with parameters (n, s, λ, α). This approximation $\tilde{\mathbf{H}}$ of $\mathbf{H}$ allows the number of strict local maxima in G_α to be estimated, Sec. 4.4.4, and therefore the probability of the WJ algorithm restarting. Sec. 4.5.1 introduces a number of graph search algorithms from the literature. In the remainder of Sec. 4.5, the cost to find a maximum degree vertex using SAWJ and WJ under both the unit and linear cost models is measured, for both AER graphs and "real-world" SNAP graphs, against several graph search algorithms from the literature.

Prop. 11 and Prop. 13 in Chap. 5 provide estimates for the number of samples required to find

a degree k vertex or a degree (j,k) edge in an ER graphs G_ϵ under SS-R sampling. It is shown in Chap. 6 that these estimates are unnecessarily complicated. The important contribution of Chap. 5 then is not the estimates in Prop. 11 and Prop. 13, but rather the justification given for $U_{j,k} \approx Y_{j,k}$, where $U_{j,k}$ it the probability that a degree j vertex in an ER graph G_ϵ has no degree k neighbors and $Y_{j,k}$ is a combinatorial function, Def. 16. This justification relies on the introduction of Modified Erdős Rényi (MER) graphs in Sec. 5.3. In a MER graph, $\tilde{G}_\epsilon$, edges are placed independently producing a multigraph where multiple edges may lie between any two vertices. In MER graphs Thm. 12 states that the approximation $U_{j,k} \approx Y_{j,k}$ holds. Thm. 12 is justified for ER graphs G_ϵ by arguing for small s and large n the number of multiple edges in $\tilde{G}_\epsilon$ goes to zero, Thm. 10, and the degree expectation and variance of graphs G_ϵ and $\tilde{G}_\epsilon$ are approximately equal, Thm. 11. Therefore given the similarity between G_ϵ and $\tilde{G}_\epsilon$ up to the second moment of their degree distributions and Thm. 12 it follows that the approximation $U_{j,k} \approx Y_{j,k}$ holds for small s and large n in ER graphs G_ϵ.

Chap. 6 accomplishes what Chap. 5 attempted, giving analytical estimates for the expected cost of finding a target set $\mathcal{V}^*$ using SS-R, SS-C, and SS-S under unit and linear cost models. In particular, Sec. 6.3 gives estimates of the expected unit costs of SS-R, SS-C, and SS-S in ER graphs and proves conditions under which asymptotically the expected costs of the three star sampling variants are equivalent. While Sec. 6.5 gives the estimates of the expected linear cost of SS-R, SS-C, and SS-S on ER graphs. The cost estimates for SS-R and SS-C sampling are relatively straight forward, however the unit and linear cost estimates for SS-S require estimating the expected number of vertices in the graph and the extended target set as a function of the number of SS-S samples taken. Numerical results for the unit and linear cost estimates for SS-R, SS-C, and SS-S on Erdős Rényi and "real-world" graphs are given in Sec. 6.6.

Having studied the effectiveness of using Star Sampling (SS) to find a maximum degree vertex $v \in \mathcal{V}_\lambda$ in Chaps. 5 and 6, Chap. 7 turns back to the rewiring algorithms used in Chaps. 2 to 4. Specifically, Chap. 7's seeks to study the use of greedy graph wiring and graph rewiring algorithms to construct graphs of maximum assortativity, over all graphs with a specified degree sequence. We

recall Taylor [1981] proved that a sequence of edge pair rewirings can traverse the set of simple connected graphs with any degree sequence $\mathbf{d}$ satisfying the Erdős Gallai theorem, Erdős and Gallai [1960]. Chap. 7 suggests efficient algorithms may not exist, as Kincaid et al. [2016] argues that the problem is NP-Hard and it is shown in Sec. 7.2 that greedy rewiring heuristics and in Sec. 7.3 that a wiring heuristic available in the literature can fail to return maximally assortative graphs.

Chap. 8 considers the task of estimating the expected number of samples without replacement required to find a near maximum value from a list of independent and identically distributed random variables if the maximum value in the list is unknown. Although this problem is not related to graphs per se, the problem of finding a maximum degree node in a graph by observing the degrees of vertices gathered by sampling, studied in previous chapters, is a special case of this problem where in addition to node degrees one can access the edge structure of the sampled graph. Sec. 8.3 applies extreme value theory (EVT) to estimate the expected maximum value in cases where the values are normally or Pareto distributed. Sec. 8.4 uses these estimates to approximate the expected number of samples required to find a near maximum value if the list of values have either a binomial or Zipf/Zeta distribution. Results for the estimates of the expected number of samples and the expected maximum value sampled using these estimates are given in the case where the parameters of the binomial and Zipf/Zeta distributions are known a priori (Sec. 8.4.3) and the case where they are estimated online (Sec. 8.4.4).

1.3 Notation

1.3.1 General Math Notation

$\mathbf{1}_x$	vector of ones of size x
$a \equiv b$	a and b are equal by def.
$x \approx y$	x is approximately equal to y
$\bar{\alpha}$, $\bar{x}$, etc.	denotes $1-\alpha$, $1-x$, the complement of α, x
x^U, X^U	upper bound on x, X
x^L, X^L	lower bound on x, X
bold lower case char.	a vector $\mathbf{m} = (m_s, s \in \mathcal{D})$
bold upper case char.	a matrix $\mathbf{M} = (M_{s,t}, (s,t) \in \mathcal{D}^2)$ or $\mathbf{M}_X = (M_X(s,t), (s,t) \in \mathcal{D}^2)$
calligraphic font	sets ie $\mathcal{X}$, $\mathcal{N}$
$\text{Bin}(\cdot,\cdot)$	binomial distribution
$\text{Corr}(\cdot,\cdot)$	correlation
$\text{Cov}(\cdot,\cdot)$	covariance
$\boldsymbol{C}_n$ (bi-var)	cov. matrix of the random endpoint degrees (X,Y) of a random edge
$\epsilon \in [0,1)$	arbitrary small number
η_1, η_2	mean of Z_1 and Z_2
$\mathbb{E}[\cdot]$	expectation
$\text{Exp}(\cdot)$, e	exponential function
$\hat{\gamma}$	Pearson correlation coefficient of Z_1 and Z_2
$\text{Geo}(\cdot)$	Geometric distribution single parameter
$\mathbf{I}, \mathbf{I}_x$	identity matrix, identity matrix's of size $x \times x$
$\boldsymbol{\kappa}_n$ (bi-var)	vector means of the random endpoint degrees (X,Y) of a random edge
$M \equiv \binom{n}{2}$	ways of choosing pairs from a set of size n
$\text{Mult}(\cdot,\cdot)$	Multinomial distribution
N_k	Multinomial distributed entries of N
$\mathsf{N} = (N_k, k \in \mathcal{D})$	random ω-vector has a multinomial distribution $\mathsf{N} \sim \text{Mult}(j, \mathbf{p})$
n, N	vectors of random variables
$\mathcal{N}(\mu_0, \sigma_0^2)$	Normal RV with parameters mean μ_0 and variance σ_0^2
$[n]$	set $[n] \equiv \{0, \ldots, n\}$ for $n \in \mathbb{N}$
$[n]^+$	set $[n] \equiv \{1, \ldots, n\}$ for $n \in \mathbb{N}$
$\mathcal{N}_k$	set of all ω-vectors, $\ddot{\mathbf{n}}$, from $\mathbb{N}^\omega$ that sum to k is the support of N
$\text{Par}(y_0, \alpha)$	Pareto distribution with scale y_0 and shape α
$\text{Poi}(\cdot)$	Poisson distribution
$\phi(t)$, $\phi_Y(t)$	moment generating function $\phi(t) \equiv \mathbb{E}[e^{tY}]$
$\mathbb{P}(\cdot)$	probability
$Q(\cdot)$	standard Normal CDF
sans-serif	ex. X, x, $\mathsf{\lambda}$, random variables
script font	collections of subsets ex. $\mathscr{E} = (\mathcal{E}_s, s \in \mathcal{D})$, $\mathscr{V} = (\mathcal{V}_{j,k}, j \in \mathcal{D}, k \in \mathcal{D})$
σ_1, σ_2	variance of random variables Z_1 and Z_2
$\text{Std}(\cdot)$	standard deviation
$\text{Uni}(\cdot)$	Uniform distribution
$\text{Var}(\cdot)$	variance
$\varphi(z)$, $\varphi_Y(z)$	probability generating function $\varphi(z) \equiv \mathbb{E}[z^Y]$
$W(\cdot)$	Lambert W function
$\mathsf{X} \approx \mathsf{Y}$	RV's X and Y have approximately equal dist.
$\mathbf{X}^\mathsf{T}$	transpose of $\mathbf{X}$
$\{\mathsf{X}\|\mathsf{Y}\}$	X conditioned on Y
$\{\mathbf{X}_n\}$	a sequence of k-dim. Bin. random vectors of mean $\boldsymbol{\kappa}_n$ and cov. $\boldsymbol{C}_n$
$(\tilde{\mathsf{X}}, \tilde{\mathsf{Y}})$	two Normal RV $(\tilde{\mathsf{X}}, \tilde{\mathsf{Y}}) \sim \mathcal{N}(\boldsymbol{\kappa}_n, \boldsymbol{C}_n)$
$(\mathsf{Z}_1, \mathsf{Z}_2)$	two bivariate Normal RV
$\zeta(c)$	Riemann zeta function

$\zeta'(c)$	derivative of the Riemann zeta function

1.3.2 Graph Notation

$\alpha = \alpha(G)$	degree assortativity of graph G
α_i	degree assortativity of G_i after the i_{th} rewiring iteration, $\alpha_i = \alpha(G_i)$
α_T	target degree assortativity
$A_{j,k}$	$A_{j,k} = A_{k,j} = \#\{e \in \mathcal{E} \vert \hat{d}_v(e) = \{j,k\}\}$ where $\{j,k\}$ are excess degrees
$A^{(1)}_{k\vert j}$	approximate expected number of degree k 1-hop neighbors of a degree j node
$A^{(2)}_{k\vert j}$	approximate expected number of degree k 2-hop neighbors of a degree j node
$\mathcal{A}(\mathbf{d})$	set of possible α's for degree sequence $\mathbf{d}$
b_k	is the probability an edge endpoints excess degree is k
$\mathbf{b} = (b_k, k \in \mathcal{B})$	excess degree distribution
$\mathbf{B} = \{B_{j,k}\}$	joint excess degree distribution with entries $B_{j,k} = B_{k,j}$
$\mathcal{B}$	set of excess degrees
$c^{\text{SS-R}}_{\max}$, $c^{\text{SS-C}}_{\max}$, $c^{\text{SS-S}}_{\max}$	upper bound on $\mathsf{c}^{\text{SS-R}}_u$, $\mathsf{c}^{\text{SS-C}}_u$, $\mathsf{c}^{\text{SS-S}}_u$
$\mathsf{c}^{\text{SS-R}}_l$, $\mathsf{c}^{\text{SS-C}}_l$, $\mathsf{c}^{\text{SS-S}}_l$	cost of SS-R, SS-C, and SS-S sampling under the linear cost model
$\mathsf{c}^{\text{SS-R}}_u$, $\mathsf{c}^{\text{SS-C}}_u$, $\mathsf{c}^{\text{SS-S}}_u$	cost of SS-R, SS-C, and SS-S sampling under the unit cost model
$C_{j,k}$	entries of $\mathbf{C}$
$\mathbf{C}_{j\cdot}$	row j of $\mathbf{C}$
$\mathbf{C}$	JDD conditioned one edge endpoint degree being fixed s.t. $\mathbf{C}$'s rows sum to 1
$\mathcal{C}_t$	event that star sample t misses the target set
$\bar{\mathcal{C}}_t$	event that the first t star samples miss the target set
d_v, $d(v)$	$d_v = \vert\Gamma_v\vert$ degree of node v
d^e_v	extended degree of node v
$d(e)$, $d(vu)$	$d(e) = d(vu) = \{j,k\}$, degrees at either end of edge e
$\tilde{d}_v$	number of wired stubs of node v in wired graph $\tilde{G}$
d^-_v	degree of node v before wiring
d^+_v	degree of node v after wiring
$d_{\mathcal{B}}$	the sum of all target degrees in the graph under wiring in set $\mathcal{B}$
$\hat{d}_v$	the excess degree of node v
$\mathbf{d}$	degree sequence of a graph, $\mathbf{d} = (d_v, v \in \mathcal{V})$
$\mathbf{d}_G$	degree sequence of graph G, $\mathbf{d}_G = (d_v, v \in \mathcal{V})$ where $G = (\mathcal{V}, \mathcal{E})$
$\mathsf{d}^{\max}$	random maximum degree of a graph
$\mathcal{D}$	$\mathcal{D} = \cup_{v\in\mathcal{V}} d_v$ set of degrees in the graph
$\bar{\mathcal{D}}$	set of NI degrees
$\tilde{\mathcal{D}}$	$\tilde{\mathcal{D}} = \{0, \ldots, \tilde{\xi}\}$ approx. state space on AER
$\mathcal{D}_v$	degree neighborhood of v
$\ddot{\mathcal{D}}_k$	non-zero entries of $\mathbf{H}_k$ and $\mathbf{F}_k$, ie degrees that are neighbors of degree k
$\mathscr{D}^{(n)}$	collection of degree sequences of $\mathcal{W}^{(n)}$
δ_j	stubs available for wiring in $\tilde{G}$, $\delta_j = d_j - \tilde{d}_j$
Δ_1, Δ_2	changes in graph assortativity after rewiring
$\Delta_{G,G'}$	the change in the s-metric due to rewiring G to G'
e_j	random edge j where j was chosen uniformly at random from $j \in [m]$
$e = uv = vu = \{uv\} = \{vu\}$	undirected edge between nodes u and v
$e = (uv)$	directed edge from node u to node v
$E_{j,k}$	$E_{j,k} \equiv \vert U_{j,k} - Y_{j,k}\vert$
$\mathbf{E} = \{E_{j,k}\}$	deviation matrix
$\tilde{E}_{j,k}$	$\tilde{E}_{j,k} = \tilde{E}_{k,j} \equiv \vert\tilde{U}_{j,k} - \tilde{Y}_{j,k}\vert$ error in $\tilde{Y}_{j,k}$'s approximation of $\tilde{U}_{j,k}$
$\tilde{E}^U_{j,k}$	upper bound on $\tilde{E}_{j,k}$
$\tilde{E}^L_{j,k}$	lower bound on $\tilde{E}_{j,k}$
ϵ_k	upper bound on the error in approximation y_k of u_k, $\vert u_k - y_k\vert \leq \epsilon_k$

$\mathcal{E}$	set of edges in a graph
$\tilde{\mathcal{E}}$	edge set of of a modified ER graph, $\tilde{G}_\epsilon$
$\hat{\mathcal{E}}$	edge set of meta graph $\hat{\mathcal{G}}$
$\ddot{\mathcal{E}}$	edge set of the graph in the wiring algorithm
$\mathcal{E}_k$	set of edges with at least one endpoint is degree k
$\bar{\mathcal{E}}_t$	edge set at iteration t of SS-C
$\breve{\mathcal{E}}_t$	edge set at iteration t of SS-S
$\mathcal{E}_{j,k}$	set of edges with degree endpoints of $\{j,k\}$ or $\{k,j\}$
$\hat{\mathcal{E}}_{\mathbf{d}}^{(H)}$	edge set in the meta-graph under heuristic $H \in \{A,B,C\}$
f_j	the probability a random endpoint of a randomly selected edge is degree j
$\mathbf{f} = \{f_j\}$	marginal distribution of $\mathbf{F}$, ie the row sums of $\mathbf{F}$, with entries f_j
f_{nst}	fraction of non-strict local degree maxima, $f_{nst} = \lvert\mathcal{V}_{nst}\rvert/\bar{n}$
f_{str}	fraction of strict local degree maxima, $f_{str} = \lvert\mathcal{V}_{str}\rvert/\bar{n}$
$\tilde{f}_{nst}$	approximation of f_{nst} for arbitrary graphs
$\tilde{f}_{str}$	approximation of f_{str} for arbitrary graphs
$\hat{f}_{nst}$	approximation of f_{nst} for AER graphs
$\hat{f}_{str}$	approximation of f_{str} for AER graphs
$F_{j,k}$	entries of the joint degree distribution $\mathbf{F}$
$\mathbf{F} = \mathbf{F}(G)$	joint degree distribution of G
$\mathbf{F}_k$	the k_{th} row of $\mathbf{F}$
$\mathcal{F}$	set of nodes with unwired stubs
$G = (\mathcal{V}, \mathcal{E})$	simple undirected graph with sets $\mathcal{V}$ nodes and $\mathcal{E}$ edges
G_0	initial graph in the rewiring algorithm
G_α	AER graph with parameters (n, s, α)
G_β	BA graph
$G_\epsilon = (\mathcal{V}, \mathcal{E})$	ER graph
$G_{\epsilon,t}$	ER graph at iteration t of rewiring
$\tilde{G}_\epsilon = (\mathcal{V}, \tilde{\mathcal{E}})$	modified ER graph
G_t	graph G at iteration t of rewiring
$\bar{G}_t$	graph G at iteration t of SS-C
$\breve{G}_t$	graph G at iteration t of SS-S
$G_0 \in \mathcal{W}_{\mathbf{d}}^{(n)}$	initial graph of the rewiring algorithm for degree sequence $\mathbf{d}$
$G_T \in \mathcal{W}_{\mathbf{d}}^{(n)}$	final graph of the rewiring algorithm for degree sequence $\mathbf{d}$
$G_{0,H}$	initial graph in counter example to heuristic for $H = \{A,B,C\}$
$G_{\mathbf{d},opt}^{(n)}$	if $\mathcal{G}_{\mathbf{d},opt}^{(n)}$ set is unique, graph $G_{\mathbf{d},opt}^{(n)}$ is the sets unique element
$G' = G(ik, jl)$	graph G' resulting from rewiring edges (ij, kl) in G to (ik, jl)
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	set of simple undirected graphs with n nodes and m edges
$\mathcal{G}_\alpha$	set of AER graphs
$\mathcal{G}_\beta$	set of BA graphs
$\hat{\mathcal{G}}_{\mathbf{d}}^{(n)} = (\mathcal{W}_{\mathbf{d}}^{(n)}, \hat{\mathcal{E}}_{\mathbf{d}})$	undirected meta-graph where the set of graphs $\mathcal{W}_{\mathbf{d}}^{(n)}$ are the node set of $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)}$
$\hat{\mathcal{G}}_{\mathbf{d},H}^{(n)} \equiv (\mathcal{W}_{\mathbf{d}}^{(n)}, \hat{\mathcal{E}}_{\mathbf{d}}^{(H)})$	directed meta-graph for heuristic $H \in \{A,B,C\}$
$\mathcal{G}_\epsilon \equiv G(n,s)$	set of ER graphs with parameters n nodes and s edge probability
γ	Euler-Mascheroni constant
Γ_v, $\Gamma_G(v)$	$\Gamma_v = \{u \in \mathcal{V} \lvert u, v \in \mathcal{E}\}$ neighbors of node v
Γ_v^e, $\Gamma_G^e(v)$	extended set of neighbors of node v
Γ_λ	neighborhood of the max degree set $\mathcal{V}_\lambda$
$\Gamma(\mathcal{Z})$	neighborhood of the set $\mathcal{Z}$ excluding $\mathcal{Z}$ itself
$\Gamma^c(\mathcal{Z})$, $\Gamma^e(\mathcal{Z})$	closure over $\mathcal{Z}$ and $\Gamma(\mathcal{Z})$, or the extended neighborhood of node set $\mathcal{Z}$
Γ^+	set of maximum degree neighbors of a randomly selected node
Γ_u^+	maximum degree neighbors of u, $\Gamma_u^+ = argmax_{v \in \Gamma_u} d_v$
Γ_u^-	minimum degree neighbors of u, $\Gamma_u^- = argmin_{v \in \Gamma_u} d_v$
$\Gamma_t(v)$	neighborhood of node v at iteration t of SS-R

$\Gamma_t^c(v)$	closure over v and v's neighborhood at iteration t of SS-R
$\Gamma(\Gamma(v))$	two hop neighbors of node v
$\bar{\Gamma}_t^c(v)$	closure over v and v's neighborhood at iteration t of SS-C
$\breve{\Gamma}_t^c(v)$	closure over v and v's neighborhood at iteration t of SS-S
h_k	probability a degree k node is a local max, with neighbors degree k or less
$\mathbf{h} = (h_k, k \in \mathcal{D})$	distribution of a degree k node being a local max
$h_{j,k} = h_{k,j}$	$\#\{e \in \mathbf{E} \vert d(e) = \{j,k\}\}$ number of edges with $\{j,k\}$ degree endpoints
$\mathsf{h}_{k\vert j}^{(1)}$	number of degree k 1-hop neighbors of a degree j node
$\mathsf{h}_{k\vert j}^{(2)}$	number of degree k 2-hop neighbors of a degree j node
$\hbar_i$	change in node i's degree due to wiring, $\hbar_i = d_i^+ - d_i^-$
$\tilde{H}_{j,k}$	approximation of $H_{j,k}$ defined as $\tilde{H}_{j,k} \equiv Q_{j,k} - Q_{j,k-1}$
$H_{j,k}$	$H_{j,k} = \frac{\mathcal{V}_{j,k}}{\vert\mathcal{V}_j\vert}$ for $(j,k) \in \bar{\mathcal{D}}$, prob. a NI deg. j node's max neighboring deg. is k
H	greedy rewiring heuristic $H \in \{A, B, C\}$
$\mathbf{H} = (H_{j,k})$	conditional maximum neighbor degree distribution
$\mathbf{H}_j$	j_{th} row of matrix $\mathbf{H}$
$\tilde{\mathbf{H}} = (\tilde{H}_{j,k})$	approximate $\mathbf{H}$ distribution for AER graph
(i,j)	directed edge
$K_{m,n}$	bipartite graph of degree m and n or $2K - series$
$K_{j,k} = K_{k,j}$	$\#\{e \in \mathcal{E} \vert d(e) = \{j,k\}\}$, count of edges with endpoints degree $\{j,k\}$
$\mathbf{K}$	joint degree matrix of the graph
$\mathsf{K}(n)$, K_n	number of analytical maximizers in random vector $\boldsymbol{\mathsf{X}}$
$\hat{\mathsf{K}}(n) = \vert\hat{\mathcal{K}}\vert$	number of empirical maximum degree nodes in Bin$(\cdot)$ distributed set of degrees
$\tilde{\mathsf{K}}(n) = \vert\tilde{\mathcal{K}}\vert$	number of empirical maximum degree nodes in Poi$(\cdot)$ distributed set of degrees
$\hat{\mathcal{K}}$	set of maximum degree nodes in Bin$(\cdot)$ distributed set of degrees
$\tilde{\mathcal{K}}$	set of maximum degree nodes in Poi$(\cdot)$ distributed set of degrees
L_M	number of local maxima (NLM)
L_S	number of strict local maxima (SLM)
$\lambda = \max(\mathcal{D})$	maximum degree of the graph G
λ_ϵ	maximum degree of an ER graph G_ϵ
λ_u	maximum degree neighbor of node u, assuming only 1 such node
$\tilde{\lambda}_\epsilon$	maximum degree node in ER graph $\tilde{G}_\epsilon$
$\boldsymbol{\lambda}_u$	maximum degree neighbors of random node $u \in \mathcal{V}$, assuming only 1 such node
$\bar{\mathcal{L}} = \{L\}$	set of leaves L in the outcome space for SS-C
$\breve{\mathcal{L}} = \{L\}$	set of leaves L in the outcome space for SS-S
m	$m = \vert\mathcal{E}\vert$ number of edges in set $\mathcal{E}$
m_k	$m_k = \vert\mathcal{E}_k\vert$ number of edge endpoints of degree k
$m_{k,k}$	$\#\{uv \in \mathcal{E}, d_v = d_u\}$ number of edges with endpoints of the same degree
$m_{j,k}$	$m_{j,k} = \vert\mathcal{E}_{j,k}\vert$ number of edges with endpoints of degree j and k
$\mathsf{m} \sim \text{Bin}(\text{M}, \text{s})$	classic ER random graph size or number of edges
$\tilde{\mathsf{m}} \sim \text{Bin}(\text{M}, \text{s})$	modified ER random graph size or number of edges
$\hat{\mathsf{m}}$	nonempty random graph size or number of edges
$\bar{M}_L$	number of leaves of type $(\bar{N}_L, \bar{P}_L)$
$\breve{M}_L$	number of leaves of type $(\breve{N}_L, \breve{P}_L)$
M	maximum neighboring degree of a randomly selected node
$\hat{\mathsf{M}}(n)$	maximum degree node of an order n graph, if that node is unique
$\mathbf{m}$	$\mathbf{m} = (m_k, k \in \mathcal{D})$ the degree k endpoint or stub counts
$M = \binom{n}{2}$	all possible edge sites in an undirected graph
$\mathcal{M}$	set of *pedges* with the largest endpoint degree product
μ	mean of the degree distribution $\mathbf{w}$
$\mu_{\mathsf{n},t}$	expected order in ER graph G_t, $\mu_{\mathsf{n},t} = \mathbb{E}[\mathsf{n}_t]$
$\mu_{\mathbf{b}}$	mean of the degree excess distribution $\mathbf{b}$
$\mu_{\mathbf{f}}$	mean of the marginal of the joint degree distribution $\mathbf{f}$

$\bar{\mu}$	mean to the degree distribution $\bar{\mathbf{w}}$ for NI nodes
$\breve{\mu}_t$	mean degree in $\breve{G}_t$
n	$n = \lvert\mathcal{V}\rvert$
n_k	$n_k = \lvert\mathcal{V}_k\rvert$
$n_G^{e,*}$	order of the extended target set
n_t	order of the graph after t samples
n_t^*	order of the target set after t samples
$\mathsf{n}_t^{e,*}$	order of the extended target set after t samples
$\bar{n}$	number of NI nodes
$\breve{n}_t$	number of nodes in $\breve{G}_t$
$\breve{n}_{k,t}$	number of degree k nodes in $\breve{G}_t$
$\hat{\mathbf{n}}_{k,0}$	estimate of $\mathbf{n}_{k,0}$, the number of degree k nodes at time $t = 0$
$\mathbf{n} = (n_k, k \in \mathcal{D})$	vector collecting the number of nodes of each degree
$\ddot{\mathsf{n}}_{k\lvert v}$	$\ddot{\mathsf{n}}_{k\lvert v} = \#\{u \in \Gamma_v \lvert d_u = k\}$ number of degree k nodes in v's neighborhood
$\ddot{\mathbf{n}}_v$	$\ddot{\mathbf{n}}_v = (\ddot{\mathsf{n}}_{k\lvert v}, k \in \mathcal{D}_v)$
$\bar{N}_L$	length of path from leaf L to root node in the SS-C sampling space
$\breve{N}_L$	length of path from leaf L to root node in the SS-S sampling space
N	samples required to hit target set in SS-R
$\bar{\mathsf{N}}$	samples required to hit target set in SS-C
$\breve{\mathsf{N}}$	samples required to hit target set in SS-S
$\mathcal{N}_v$	edge neighborhood of node v
$\mathcal{N}_v^e$	extended edge neighborhood of node v
$\mathcal{N}_t(v)$	edge neighborhood of node v at time t
$\mathcal{N}_{\mathbf{d},G}$	neighborhood of G in $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)}$
$\mathcal{N}_{\mathbf{d},G}^{(H)}$	neighborhood of G in $\hat{\mathcal{G}}_{\mathbf{d},H}^{(n)}$ under rewiring heuristic $H \in \{A, B, C\}$
ν	expected degree of a randomly selected stub
ν_t	mean stub degree of a randomly selected stub at iteration t
ω	$\omega = \lvert\mathcal{D}\rvert$ is the number of distinct degrees in a graph
$\mathcal{O}$	set of *pedges* in wiring algorithm
p_k	probability a degree k node is in the neighborhood of a degree j node
p_t	prob. of sampling $\mathcal{Z}$ first at iteration t in SS-R
p_0	$\mathbb{P}(\tilde{\mathsf{m}} = 0)$, probability of zero edges in a modified ER graph
$\hat{p}_0$	$1/\bar{p}_0 = 1/(1 - p_0)$
$\bar{\mathsf{p}}_t$	probability of sampling set $\mathcal{Z}$ for the first time at iteration t in SS-C
$\breve{\mathsf{p}}_t$	probability of sampling set $\mathcal{Z}$ for the first time at iteration t in SS-S
$\breve{p}_t \equiv \mathbb{E}[\breve{\mathsf{p}}_t]$	$\breve{p}_t$ is the approximate expected value of $\breve{\mathsf{p}}_t$
$p_{k\lvert j}^{(1)}(h)$	$p_{k\lvert j}^{(1)}(h) \approx \mathbb{P}(\mathsf{h}_{k\lvert j}^{(1)} = h)$ probability a deg. j node neighbors h deg. k nodes
pedge	potential edge
$\mathbf{p} = \{p_k\}$	vector of $\mathbf{p} = (p_k, k \in \mathcal{D})$
ϕ	diameter of graph G
$\bar{\mathbf{p}}_{\mathcal{L}} \equiv (\bar{P}_L, L \in \bar{\mathcal{L}})$	probability dist. for reaching a leaf $L \in \bar{\mathcal{L}}$ under SS-C sampling
$\breve{\mathbf{p}}_{\mathcal{L}} \equiv (\breve{P}_L, L \in \breve{\mathcal{L}})$	probability dist. for reaching a leaf $L \in \tilde{\mathcal{L}}$ under SS-S sampling
$\mathcal{P}$	set of vertex property values
$\breve{\mathsf{P}}_t$	unconditioned probability of sampling $\mathcal{Z}$ in SS-S for the first time at t
$\breve{P}_t \equiv \mathbb{E}[\breve{\mathsf{P}}_t]$	$\breve{P}_t$ is the approximate expected value of $\breve{\mathsf{P}}_t$
$Q_{j,k}$	a Normal CDF $Q(f(j,k))$ raised to j which is a term in $\tilde{H}_{j,k} \approx H_{j,k}$
$\mathcal{Q}$	set of nodes with NO edges in the α wiring algorithm
$r_1(\mathcal{E}), r_2(\mathcal{E})$	two possible edges rewiring of a pair of edges (u_1, v_1), (u_2, v_2)
$r_k = w_{k-1}/w_k > 0$	entries of $\mathbf{r}$
$\mathbf{r} = (r_1, \ldots, r_{n-1})$	vector PMF ratio
ρ	simple graph undirected edge density $\rho = 2m/((n-1)n)$
$R_k = W_k/W_{k-1} > 1$	entries of vector $\mathbf{R}$

$\mathbf{R} = (R_1, \ldots, R_{n-1})$	vector CDF ratios (exception to vector notation)
$\mathcal{R}$	set of nodes with one or more edges in α wiring algorithm
s, $s(n)$	edge probability between a pair of nodes in as ER graph
s_k	$s_k \equiv \lvert\mathcal{S}_k\rvert$, number of stubs tied to degree k nodes
$s(G)$	s-metric for graph G
$\mathbf{s} \equiv (s_k, k \in \mathcal{D})$	vector of stub degree counts
$S_{\mathcal{R}}$	sum of all current degrees in the graph under wiring for nodes in set $\mathcal{R}$
$\mathcal{S}_k \subseteq \mathcal{S}$	set of stubs that are tied to degree k nodes
$\mathcal{S} = [2m]$	set of stubs in G
σ^2	variance of degree distribution $\mathbf{w}$
$\sigma^2_{\mathsf{n},t}$	variance in the order in ER graph G_t, $\sigma^2_{\mathsf{n},t} = \mathrm{Var}(\mathsf{n}_t)$
$\sigma^2_{\mathbf{b}}$	variance of excess degree distribution $\mathbf{b}$
$\sigma^2_{\mathbf{f}}$	variance of the marginal of the joint degree distribution $\mathbf{f}$
θ_j	degree of stub j
Θ_{u_k, y_k}	relative error between u_k and its approximation y_k
$\Theta_{\tilde{U}_{j,k}, \tilde{Y}_{j,k}}$	relative error between u_k and its approximation y_k
u_k	probability a randomly selected node has no degree k neighbors
u_j	$\mathsf{u}_j = \mathrm{Uni}(\mathcal{V}_j)$, a randomly selected node of degree j
$U_{j,k}$	probability a degree j node has no degree k nodes in its neighborhood
$\tilde{U}_{j,k}$	probability a degree $\{j,k\}$ edge is in the neighborhood of a random node
$\mathbf{U} = \{U_{j,k}\}$	matrix with entires $U_{j,k}$
$\mathcal{V}$	set of nodes, $\mathcal{V} = [n]$ or $\mathcal{V} = [n^+]$ depending on node indexing
$\mathcal{V}_\lambda$	set of maximum degree nodes
$\mathcal{V}_k$	$\mathcal{V}_k = \{v \in V \lvert d_v = k\}$ is the set of nodes of degree k
$\mathcal{V}^*$	a target set of elements or state, also denoted by $\mathcal{Z}$
$\mathcal{V}_0$	set of isolated nodes
$\bar{\mathcal{V}}$	set of NI nodes
$\mathcal{V}_{j,k}$	nodes of degree j with a maximum neighbor of degree k
$\mathcal{V}_{str}$	set of strict local degree maxima
$\mathcal{V}_{nst}$	set of non-strict local degree maxima
$\mathcal{V}_{\mathcal{Z}}$	target set $\mathcal{V}_{\mathcal{Z}} \subset \mathcal{V}$
$\bar{\mathcal{V}}_t$	node set at iteration t of SS-C
$\breve{\mathcal{V}}_t$	node set at iteration t of SS-S
w_k	$w_k = \frac{\lvert\mathcal{V}_k\rvert}{n}$
$\tilde{w}_k$	probability of selecting a degree k node in $\tilde{\mathcal{D}}$
$\breve{\mathsf{w}}_{k,t}$	fraction of degree k nodes in $\breve{\mathcal{V}}_t$
$\mathbf{w}$	$\mathbf{w} = (w_k, k \in \mathcal{D})$, degree distribution of G
$\tilde{\mathbf{w}}$	$\tilde{\mathbf{w}} = (\tilde{w}_k, k \in \tilde{\mathcal{D}})$, degree distribution of $\tilde{G}_\epsilon$
$\bar{\mathbf{w}}$	degree distribution of NI node in G
$\hat{\mathbf{w}}_{k,0}$	estimate of $\mathbf{w}_{k,0}$ the number of degree k nodes at $t = 0$
W_k	$W_k = \sum_{j=0}^{k} w_j$ the entries of the CDF
$\mathbf{W} = (W_k, k \in \langle n \rangle)$	CDF of the degree distribution
$\mathcal{W}^{(n)}$	all simple connected graphs of order n
$\mathcal{W}^{(n)}_{\mathbf{d}}$	a degree class, ie all simple connected graphs with degree sequence $\mathbf{d}$
$\mathcal{W}^{(n)}_{\mathbf{d},opt}$	set of graphs that achieve the maximum assortativity over $\mathcal{W}^{(n)}_{\mathbf{d}}$
x_{max}	maximum of value $\{\mathsf{x}_1, \ldots, \mathsf{x}_n\}$
X_v	random degree of node v
$\hat{\mathsf{X}}_v = d_v$	random degree of node v for an ER graph $\sim \mathrm{Bin}(\mathrm{n}-1, \mathrm{s})$
$\mathbf{X} = (\mathsf{X}_i, i \in [n])$	random vector of node degrees for G_ϵ falsely assuming degree independence
$\tilde{\mathbf{X}} = (\tilde{\mathsf{X}}_1, \ldots, \tilde{\mathsf{X}}_M)$	random vector of the occupancy counts for the $M = \binom{n}{2}$ possible edge sites
$\hat{\mathbf{X}} = (\hat{\mathsf{X}}_i, i \in [n])$	random vector of node degrees for an ER graph
y_k	approximate of u_k prob. a randomly selected node has no degree k neighbors

$y(k,n)$	upper bound on the number of maximizers K_n in a binomial graph
$\tilde{y}^{(n)}$	fraction of non-multiple edges in a modified ER $\tilde{G}_\epsilon$
$\tilde{y}_i \vert \hat{\mathsf{m}}$	RV indicating edge site i holds exactly one edge
Y_{max}	maximum of RV set $\{\mathsf{Y}_1, \ldots, \mathsf{Y}_n\}$
Y	$Y = \sum_{k \in \mathcal{D}} Y_k$, the approx. change in n given a SS-S sample
Y_k	approx. change in n_k given a SS-S sample
$Y_{k\vert j}$	approx. change in n_k given a SS-S sample centered on a degree j node
$Y^{(1)}_{k\vert j}$	approx. change in n_k if 1-hop neighbors of a random $v \in \mathcal{V}_j$ are removed
$Y^{(2)}_{k\vert j}$	approx. change in n_k if 2-hop neighbors of a random $v \in \mathcal{V}_j$ are removed
$\breve{\mathsf{Y}}_t$	change in number of nodes between $\breve{G}_{t-1}$ and $\breve{G}_t$ in SS-S
$\breve{\mathsf{Y}}_{k,t}$	change in number of degree k nodes between $\breve{G}_{t-1}$ and $\breve{G}_t$ in SS-S
$Y_{j,k}$	an approximation of $U_{j,k}$
$\tilde{Y}_{j,k}$	an approximation of $\tilde{U}_{j,k}$
$\mathbf{Y} = \{Y_{j,k}\}$	an approximation of $\mathbf{U}$
$\hat{\mathbf{Y}} = (\hat{Y}_{k,t})$	$\hat{Y}_{k,t}$ estimates $\breve{\mathsf{Y}}_{k,t}$ the change in degree k nodes at t under SS-S
z_k	$z_k = s_k/(2m)$ probability of selecting a random stub of degree k
z	sample size
$\mathbf{z} \equiv (z_k, k \in \mathcal{D})$	stub degree distribution
$\hat{\mathbf{z}}_{k,0}$	estimate of $\mathbf{z}_{k,0}$
z	degree of a random node selected from G_ϵ
$\tilde{\mathsf{z}}$	degree of a random node selected from $\tilde{G}_\epsilon$
$\tilde{\mathsf{z}} \vert \tilde{\mathsf{m}}$	conditional random degree of an arbitrary vertex

1.3.3 DTMC and BRW Notation

$\mathbf{A}$	$\hat{n} \times \hat{n}$ submatrix of the transition probabilities between absorbing states
$\mathcal{A}$	absorbing states of (DTMC)
β, $\beta(\alpha)$	BRW bias parameter or SAWJ jump bias, BRW bias as a function of α
$[\beta_-(\alpha_t), \beta_+(\alpha_t)]$	β interval where the absorption time is invariant containing β^*
$\beta^*(\alpha_t)$	optimal bias as a function of α_t
$\hat{\mathcal{D}} = \{d_\lambda\}$	absorbing states of $\mathcal{D}$
$\breve{\mathcal{D}} = \mathcal{D} \backslash \hat{\mathcal{D}}$	transient states of $\mathcal{D}$
$\boldsymbol{\mu}_U$	mean of the absorption time starting at state u
$\mathbf{N}_U$	$\breve{n} \times \breve{n}$ fundamental matrix of DTMC U, $\mathbf{N}_U = (\mathbf{I}_{\breve{n}} - \mathbf{S})^{-1}$
$\mathbf{O}$	$\hat{n} \times \breve{n}$ submatrix of the zero transition probabilities
p_{uv}, $\rho(u,v)$	BRW transition probability on an undirected graph
$\breve{p}_u$	probability of starting in transient state u
$\breve{p}_l(\ddot{\mathbf{n}}_v)$	BRW transition prob. to degree l node from node v
$\breve{p}_l(\ddot{\mathbf{n}})$	BRW transition prob. to degree l node from a node with DN $\ddot{\mathbf{n}}$
$\breve{\mathbf{p}}(\ddot{\mathbf{n}})$	BRW degree transition probability distribution
$\breve{\mathbf{p}}_U = (\breve{p}_u, u \in \breve{\mathcal{U}})$	probability distribution of starting in state u, ie $U(0) \sim \breve{\mathbf{p}}_U$
$\mathbf{P}$	$n \times n$ transition matrix
$\bar{\mathbf{P}}$	average BRW degree transition matrix size $\lambda \times \lambda$
$\tilde{\mathbf{P}}$	approximate BRW degree transition matrix
$\mathbf{P}_U$	$n \times n$ transition matrix on state space U
$\mathbf{P}_U(u,v)$,	$\mathbb{P}(U(t+1) = v \vert U(t) = u)$ the entries of $\mathbf{P}_U$
l_u	probability of starting in transient or absorbing state u
$\mathbf{l}$	probability distribution of starting state $\mathbf{l} \equiv (l_u, u \in \mathcal{U})$
$L_{s,t}$	entries of $\mathbf{L}$
$\mathbf{L}$	$\breve{n} \times \breve{n}$ submatrix of the transition probabilities between transient states
$\mathbf{M}$	$\breve{n} \times \hat{n}$ submatrix of the transition prob. from transient to absorbing states
$\boldsymbol{\sigma}^2_U$	variance of the absorption time in DTMC U

T, T_U	time to absorption, time to absorption on state space U
$\mathsf{T}_U(u)$	$T_U(u) = min\{t \in \mathbb{N} \vert U(t) \in \hat{\mathcal{U}} \vert U(0) = u\}$ absorption time from state u
$\mathsf{T}(\alpha_T, \beta)$	absorption times given α_T, β on G or a DTMC on degree states
$\mathsf{T}^*(\alpha_T, \beta^*)$	optimal absorption times given α_T, β^* on G or DTMC of degrees states
$\mathbf{T}_U = (\mathsf{T}_U(u), u \in \check{\mathcal{U}})$	collection of $\check{n}$ random absorption times for each transient state
$\boldsymbol{\tau}$	$\boldsymbol{\tau} = \mathbb{E}[\mathsf{T}_U]$ mean time to absorption
$\boldsymbol{\tau}_u$	mean time to absorption conditioned on starting at state u
$\mathscr{T}_U$	collection of absorbing times for state space U
$U = (U(t), t \in \mathbb{N})$	finite-state discrete-time Markov chain (DTMC) U
$\mathcal{U}$	finite state space
$\hat{\mathcal{U}}$	set of absorbing states
$\check{\mathcal{U}}$	set of transient states
$\vert\mathcal{U}\vert = n$	size of the finite state space
$\vert\hat{\mathcal{U}}\vert = \hat{n}$	size of the set of absorbing states
$\vert\check{\mathcal{U}}\vert = \check{n}$	size of the set of transient states
$V = (V(t), t \in \mathbb{N})$	absorbing DTMC on G with state space $\mathcal{V}$ and bias $\beta \geq 0$
$Z = (Z(t), t \in \mathbb{N})$	approximate absorbing DTMC on state space $\mathcal{D}$ with bias $\beta \geq 0$
$\mathcal{Z}$	a target set of elements or state, also denoted by $\mathcal{V}^*$

1.3.4 Set Sampling Notation

$\hat{a}$	estimate of the Pareto exponent
$\hat{c}$	estimate of the Zipf/Zeta exponent
$\delta \in [0, 1)$	the "nearness" parameter
$f_{\mathsf{X}}(x)$	PDF of X
$F_{\mathsf{X}}(x)$	CDF of X
$\gamma_{\mathsf{X}}(k, n)$	$\mathrm{Cov}(\mathsf{X}_{k:k}, \mathsf{X}_{n:n})$, covariance of max over k, n values
$\gamma_{\mathsf{X}_1,\mathsf{X}_2}$	$\mathrm{Cov}(\mathsf{X}_1, \mathsf{X}_2)$, covariance
$I(\delta)$	(unknown) set of indices with near-maximum values
$k(\delta), \tilde{k}(\delta)$	required and approximate required sample size
k_{min}	minimum number of samples to form an estimate
$k(\pi)$	minimum samples before satisfying $x_i/x_{max} \geq 1 - \epsilon$
$k_X(\delta)$	mean sample size to find an index with value in $I(\delta)$
K	represents $k(\pi)$ if π is selected uniformly at random
$\kappa(\delta)$	required fraction of the sample size $k(\delta)$ to n
$\tilde{\kappa}(\delta)$	required fraction of the approximate sample size $\tilde{k}(\delta)$ to n
$\hat{\mu}_0$	estimate of a mean of Normal RV
μ_{B}	expected maximum value of a binomial distribution
μ_{A}	expected maximum value of a Zipf/Zeta distribution
μ_{X}	$\mathbb{E}[\mathsf{X}]$ expected value of RV X
$\mu_{\mathsf{X}}(k)$	$\mathbb{E}[\mathsf{X}_{k:k}]$, expectation of max over k values
$\tilde{\mu}_{\mathsf{Z}}(n), \tilde{\sigma}^2_{\mathsf{Z}}(n)$	approximations for $\mu_{\mathsf{Z}}(n), \sigma^2_{\mathsf{Z}}(n)$
$\tilde{\mu}_{\mathsf{W}}(n), \tilde{\sigma}^2_{\mathsf{W}}(n)$	approximations for $\mu_{\mathsf{W}}(n), \sigma^2_{\mathsf{W}}(n)$
$\tilde{\mu}_{\mathsf{Y}}(n), \tilde{\sigma}^2_{\mathsf{Y}}(n)$	approximations for $\mu_{\mathsf{Y}}(n), \sigma^2_{\mathsf{Y}}(n)$
$\nu_{\mathsf{X}}(k, n)$	expected product of the maxima of two sets, $\mathbb{E}(\mathsf{X}_{k:k}\mathsf{X}_{n:n})$
p_X	$(p_X(x), x \in \mathcal{X})$ the value distribution of X
$p_X(x)$	$\#\{i \in [n] \vert x_i = x\}/n$ fraction of indices with value x
π	a permutation of $\{1, \ldots, n\}$ as $(\pi_1, \ldots, \pi_n)$
$\mathsf{query}_X(i)$	query of index $i \in [n]$ of X to learn value x_i
$r(k), \tilde{r}(k)$	expected value ratio, approximate expected value ratio
σ^2_{B}	variance of the maximum value of the binomial distribution
σ^2_{A}	variance of the maximum value of the Zipf/Zeta distribution
σ^2_{X}	variance $\mathrm{Var}(\mathsf{X})$

$\hat{\sigma}_0^2$	estimate of the variance of Normal RV
$\sigma^2_{\mathsf{X}}(k)$	$\mathrm{Var}(\mathsf{X}_{k:k})$, variance of max over k values
x_{max}	maximum element of dataset X
x_{U}	element with index U chosen uniformly at random
X	$(x_1, \ldots, x_n)$ the large unsorted dataset
$\mathsf{X}_{k:k}$	$\max_{i\in[k]} \mathsf{U}_i$, max over first k values
$(\mathsf{X}_{1:n}, \ldots, \mathsf{X}_{n:n})$	permutation of the RV's $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ in non-deceasing order
$\mathcal{X}$	$\bigcup_{i\in[n]} x_i$ the set of values in X
$\hat{y}_0$	estimate of the Pareto RV parameter
Y_∞	Frechet RV

1.3.5 Algorithms, Acronyms

AER	assortative Erdős Rényi random graphs
AER MC Model	markov chain model on the AER graphs
AS	Albatross Sampling
BA	Barabasi Albert graph
BFS	sreadth first search
CCDF	complementary CDF
CDF	cumulative distribution function
DN	degree neighborhood
ER	Erdős Rényi graph
EVT	Extreme Value Theory
FS	Frontier Sampling
IID	independent and identically distributed
JDD	joint degree distribution
JDM	joint degree matrix of G
MC Model	markov chain model
MER	modified Erdős Rényi graphs
NI	non-isolated nodes
NLM	non-strict local degree maximum
PDF	probability density function
PMF	probability mass function
RG	random graph
RS	random sampling with replacement
RSW	random sampling without replacement
SS	star sampling
SS-C	star sampling without center replace
SS-R	star sampling with replace
SS-S	star sampling without star replacement
SAWJ	self-avoiding walk-jump
SLM	strict local degree maximum
SRA	stochastic rewiring algorithm
WJ	Walk-Jump algorithm

Chapter 2: A Markov chain model for the search time for max degree nodes in a graph using a biased random walk

2.1 Introduction

A graph representing Facebook's network of 1.4 billion users would require 1.4 billion nodes and hundreds of billions of edges, stretching the capacity of current hardware to hold the graph in memory. This inability to represent large graphs makes them difficult to study. One way of sidestepping this issue is to study a representative subsample of the entire graph. How one takes this sample often depends on the properties of the graph being studied. The simplest method of sampling a graph is to select nodes uniformly at random, however, many graphs, specifically those representing social networks, exhibit a power law degree distributions, therefore the probability of selecting a high degree node using this method is small. Intuitively a better sampling method would use the information gained by a sample to increase the probability of selecting a max degree node on subsequent samples. In the context of social networks, this chapter looks at sampling methods for a max degree node that exploit local information exploit the friendship paradox; on average your friends have more friends than you do. *One goal is to study the impact that biasing the random choice of the next neighbor in a random walk towards selecting higher degree neighbors has on the time to reach a maximum degree node.*

Previous work has developed analytical bounds for the hitting time of a biased random walk (BRW), the time it takes to get from one node in a graph to another, and the cover time of the walk, the time it takes a walk to visit every node in a graph. Ikeda shows that the hitting and cover time of an undirected graph of n nodes is upper bounded by $O(n^2)$ and $O(n^2 log(n))$ respectively Ikeda and Kubo [2003]. Cooper shows all nodes of degree n^a for $a < 1$ in a n node power law graphs with exponent c can be found in $O(n^{1-a(c-2)+\delta})$ steps Cooper et al. [2014]. Maiya evaluates several biased sampling strategies on real world graphs numerically showing that a walk which always transitions to a max degree neighboring node is a good method of exploring these graphs Maiya and Berger-Wolf [2011].

This chapter looks at a more specific problem, how to efficiently find a maximum degree node in a graph of order n. Observe that the basic theory of absorbing finite-state Markov chains (Thm. 1) yields expressions for the mean and variance of the random time to reach an absorbing state. Defining the maximum degree nodes on an n-node graph as the absorbing state of a Markov chain representing

a random walk (RW) then gives mean and variance in the time for the RW to find a maximum degree node. However, *computing* this mean and variance requires inverting an $n \times n$ matrix, which for large graphs is prohibitive (Sec. 2.3).

To circumvent this issue, the approach taken in this chapter restricts itself to finding maximum degree nodes in assortative graphs where the degrees of the endpoints of an edge are positively correlated. In assortative graphs an intelligent strategy for minimizing the search time to find a maximum degree node is to exploit the local gradient provided by the assortativity, and to select the next node in the walk with a bias towards higher degree neighbors. This search strategy can be modeled as a random walk on a significantly reduced state space, with one state for each degree in the graph, and the transition probability matrix derived from the joint degree distribution of the graph and the random walk bias parameter (Sec. 2.4). The advantage of such a model is that it can be used to analytically compute the absorption time of the random walk on this reduced state space quite easily. *A second goal is to study how the statistics of the random absorption time to find maximum degree nodes on large graphs may be captured using this model.*

Specifically, as shown in Sec. 2.5, this chapter investigates three natural questions, for which it offers only preliminary and numerical results: *i*) for which graphs does the above Markov chain state reduction accurately capture the mean absorbtion time?, *ii*) for which graphs does a biased random walk find a maximum degree node more quickly than does random sampling?, and *iii*) how does the optimal bias parameter (β) depend upon the graph? In the preliminary numerical investigation of these three questions, this chapter employs Erdős-Rényi (ER) graphs, rewired (using a standard rewiring algorithm) to ensure the graph is connected and has a target assortativity ($\alpha \in [-1, +1]$); the assortativity plays the role of an independent control parameter in these investigations.

The results presented in this chapter suggest the following answers to the above questions. First, there are certain (α, β) pairs for which the above model does and does not work well, in particular $\alpha < 0$ and β large gives a poor match. Second, for certain values of α the optimal $\beta^*(\alpha)$ are such that the resulting optimal absorbtion time of a biased random walk is superior to random sampling, but otherwise not. Third, the optimal $\beta^*(\alpha)$ appears to have an increasing trend in α, so that "following the local gradient (choosing a maximum degree neighbor)" is optimal for highly assortative graphs, while choosing a neighbor uniformly is superior for highly disassortative graphs.

2.2 Absorption time for Markov chains

This chapter requires the theorem and corollary below on the mean and variance of the random absorption time T_U of an absorbing finite-state discrete-time Markov chain (DTMC) $U = (U(t), t \in$

$\mathbb{N}$) taking values in a finite state space $\mathcal{U}$ (with $|\mathcal{U}| = n$). Partition $\mathcal{U}$ into absorbing states $\hat{\mathcal{U}}$ (with $|\hat{\mathcal{U}}| = \hat{n}$) and transient states $\check{\mathcal{U}}$ (with $|\check{\mathcal{U}}| = \check{n} = n - \hat{n}$), and so too partition the $n \times n$ transition matrix $\mathbf{P}_U$ into submatrices

$$\mathbf{P}_U = \begin{array}{c} \\ \hat{u} \\ \check{u} \end{array} \begin{array}{c} \begin{array}{cc} \hat{u} & \check{u} \end{array} \\ \begin{bmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{M} & \mathbf{L} \end{bmatrix} \end{array}, \tag{2.1}$$

where $\mathbf{A}$ is the $\hat{n} \times \hat{n}$ submatrix of transition probabilities between absorbing states, $\mathbf{O}$ is the $\hat{n} \times \check{n}$ zero matrix, $\mathbf{M}$ is the $\check{n} \times \hat{n}$ submatrix of transition probabilities from transient to absorbing states, and $\mathbf{L}$ is the $\check{n} \times \check{n}$ submatrix of transition probabilities between transient states.

Definition 1. *The* fundamental matrix *for the discrete-time Markov chain (DTMC) U is the $\check{n} \times \check{n}$ matrix $\mathbf{N}_U = (\mathbf{I}_{\check{n}} - \mathbf{L})^{-1}$, for $\mathbf{I}_{\check{n}}$ the $\check{n} \times \check{n}$ identity matrix. Define $\mathscr{T}_U = (\mathsf{T}_U(u), u \in \check{\mathcal{U}})$ where $\mathsf{T}_U(u) = min\{t \in \mathbb{N}, U(t) \in \hat{\mathcal{U}} | U(0) = u\}$ as the collection of $\check{n}$ random absorption times starting from each possible initial transient state.*

Theorem 1. *(Kemeny and Snell [1983]) The absorption times $\mathscr{T}_U$ have means $\boldsymbol{\mu}_U$ and variances $\boldsymbol{\sigma}_U^2$:*

$$\boldsymbol{\mu}_X = \mathbf{N}_U \mathbf{1}_{\check{n}}, \ \boldsymbol{\sigma}_U^2 = (2\mathbf{N}_U - \mathbf{I}_{\check{n}})\boldsymbol{\mu}_U - (\boldsymbol{\mu}_U^{\mathsf{T}} \boldsymbol{\mu}_U)\mathbf{1}_{\check{n}}, \tag{2.2}$$

where $\mathbf{1}_{\check{n}}$ is the $\check{n}$-vector of ones.

Let $\check{\mathbf{p}}_U = (\check{p}_u, u \in \check{\mathcal{U}})$ be the initial distribution of U on the transient states $\check{\mathcal{U}}$, i.e., $\mathsf{U}(0) \sim \check{\mathbf{p}}_U$, and define T_U as the corresponding random absorption time.

Corollary 1. *Given the initial distribution $\check{\mathbf{p}}_U$ for $U(0)$ on $\check{\mathcal{U}}$, the resulting absorption time T_U has*

$$\mathbb{E}[\mathsf{T}_U] = \sum_{u \in \check{\mathcal{U}}} \check{\mathbf{p}}_U(u) \boldsymbol{\mu}_U(u), \ \mathrm{Var}(\mathsf{T}_U) = \sum_{u \in \check{\mathcal{U}}} \check{\mathbf{p}}_U(u) \boldsymbol{\sigma}_U^2(u). \tag{2.3}$$

2.3 Biased random walk on the graph

Denote the undirected graph on which the search takes place as $G = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ the set of vertices and $\mathcal{E}$ the set of undirected edges. The order and size of G are $|\mathcal{V}| = n$ and $|\mathcal{E}| = m$, respectively. The neighborhood of any node v is denoted $\Gamma_v = \{u \in \mathcal{V} | uv \in \mathcal{E}\}$, and the degree is $d_v = |\Gamma_v|$. The set of degrees found in the graph is $\mathcal{D} = \bigcup_{v \in \mathcal{V}} d_v$, the number of distinct degrees is $\omega = |\mathcal{D}|$. The nodes $\mathcal{V}$ are partitioned into subsets $(\mathcal{V}_k, k \in \mathcal{D})$ with $\mathcal{V}_k = \{v \in \mathcal{V} | d_v = k\}$ the set of nodes of degree k, and the resulting degree distribution is $\mathbf{w} = (w_k, k \in \mathcal{D})$, with entries $w_k = \frac{|\mathcal{V}_k|}{n}$. In

particular, the max degree found in the graph is $\lambda = \max(\mathcal{D})$, and the set of nodes with max degree is $\mathcal{V}_\lambda$. An absorbing biased random walk on G is defined as follows, where the notation is adopted from the general case presented in Thm. 1.

Definition 2. *The absorbing discrete-time Markov chain (DTMC) $V = (V(t), t \in \mathbb{N})$ on G with bias parameter $\beta \geq 0$ has states $\mathcal{V}$, partitioned into absorbing states $\hat{\mathcal{V}} = \mathcal{V}_\lambda$ and transient states $\check{\mathcal{V}} = \mathcal{V} \setminus \mathcal{V}_\lambda$. The $n \times n$ transition probability matrix $\mathbf{P}_V$ has entries $\mathbf{P}_V(u,v) = \mathbb{P}(V(t+1) = v | V(t) = u)$. For $u \in \check{\mathcal{V}}$ set $\mathbf{P}_V(u,v) = \frac{d_v^\beta}{\sum_{w \in \Gamma_u} d_w^\beta}$ for $v \in \Gamma_u$ and $\mathbf{P}_V(u,v) = 0$ else. For $u \in \hat{\mathcal{V}}$, $\mathbf{P}_V(u,u) = 1$ and $\mathbf{P}_V(u,v) = 0$ for $v \neq u$. The absorption times from each initial transient state are $\mathscr{T}_V = (\mathsf{T}_V(v), v \in \check{\mathcal{V}})$, with $\mathsf{T}_V(v) = \min\{t \in \mathbb{N}, V(t) \in \hat{\mathcal{V}} | V(0) = v\}$. Note the event $V(t) \in \hat{\mathcal{V}}$ is equivalent to $d(V(t)) = \lambda$.*

This random search for a node in $\mathcal{V}_\lambda$ uses a transition probability that is biased towards selecting higher degree neighbors of the current node, as introduced by Cooper et al. [2014], with the bias monotonically increasing in β. For $\beta = 0$ the next step is uniform among all neighbors of u, while as $\beta \to \infty$ the next step is uniform among the maximum degree neighbors of u, i.e., $\Gamma_u^+ = \text{argmax}_{v \in \Gamma_u}(d_v)$. A key goal of this chapter is to characterize $\mathbb{E}[\mathsf{T}_V]$ and $\text{Var}(\mathsf{T}_V)$ as a function of the initial distribution $\check{p}_V$ on $\check{\mathcal{V}}$ and on the bias parameter β. Although numerical estimates can (and will) be obtained by simulating the random walk, for large graphs (large n) it is computationally infeasible to compute $\mathbb{E}[\mathsf{T}_V]$ and $\text{Var}(\mathsf{T}_V)$ analytically using Cor. 1, due to the need to compute $\mathbf{N}_V$ as the inverse of the (large) $\check{n} \times \check{n}$ matrix $\mathbf{I}_{\check{n}} - \mathbf{L}$. This difficulty motivates us to define an approximation of the biased random walk using a significantly smaller state space, discussed next.

2.4 Approximate biased random walk

In this section develops the chapters key approximation giving a more computationally feasible means to estimate the mean and standard deviation of T_V. The derivation consists of the following steps: *i*) define the joint degree matrix $\mathbf{F}$ and the conditional degree distribution matrix $\mathbf{H}$ for G, *ii*) define the biased walk degree transition probability distribution $\check{\mathbf{p}}(\ddot{\mathbf{n}})$ in terms of a degree neighborhood $\ddot{\mathbf{n}}$ and average biased walk degree transition matrix $\bar{\mathbf{P}}$, and *iii*) define the approximate biased walk degree transition matrix $\tilde{\mathbf{P}}$.

2.4.1 Joint degree and conditional degree distribution matrices

The definition for the joint degree matrix $\mathbf{K}$ and the conditional degree distribution matrix $\mathbf{H}$ of an arbitrary graph G is given below.

Definition 3. *The joint degree matrix* $\mathbf{K}$ *is the* $\omega \times \omega$ *symmetric matrix with entries* $K_{j,k} = \#\{uv \in \mathcal{E} | \{d_u, d_v\} = \{j,k\}\}$*, the number of edges in* G *with endpoints of degrees* j *and* k*, and entries* $K_{j,j}$ *twice the number of edges with endpoints both of degree* j*, for* $(j,k) \in \mathcal{D}^2$*.*

Example 1. *The joint degree matrix* $\mathbf{K}$ *for the graph* G *shown in Fig. 2.1 (left) is found by grouping the edges* $\mathcal{E}$ *by the degrees of the endpoints, as shown in Fig. 2.1 (right):*

$$\mathbf{K} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} & & & 1 \\ & & 1 & 3 \\ & 1 & & 2 \\ 1 & 3 & 2 & 2 \end{bmatrix} \end{array}. \tag{2.4}$$

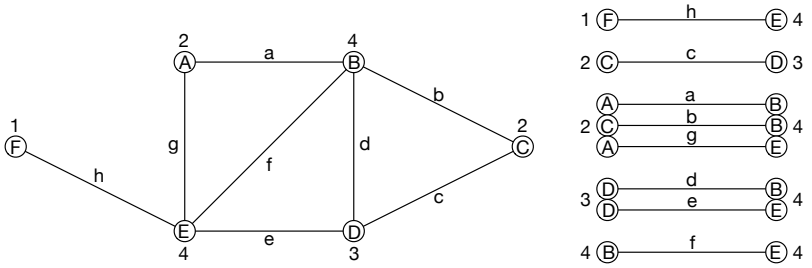

Figure 2.1: Graph G (left) with edges $\mathcal{E}$ grouped by the endpoint degrees (right).

Observe $\sum_{j,k} K_{j,k} = 2m$, and that if an edge from $\mathcal{E}$ is selected uniformly at random then the probability that it should have endpoints $\{j,k\}$ is given by $K_{j,k}/m$ for $j \neq k$, and $K_{j,j}/(2m)$ else. The conditional degree distribution matrix is obtained by normalizing each row of $\mathbf{K}$ into a probability distribution.

Definition 4. *The conditional degree distribution matrix* $\mathbf{C}$ *is the* $\omega \times \omega$ *matrix with* $C_{j,k} = K_{j,k} / \sum_{k' \in \mathcal{D}} K_{j,k'}$*, for* $(j,k) \in \mathcal{D}^2$*.*

To motivate $\mathbf{C}$ it is necessary to define the following. First, the set of degrees $\mathcal{D}_v \subseteq \mathcal{D}$ of some $v \in \mathcal{V}$ is formed by taking the union of the degrees of each neighbor of v, i.e., $\mathcal{D}_v = \bigcup_{u \in \Gamma_v} d_u$. The number of neighbors of some $v \in \mathcal{V}$ of each different degree $\mathcal{D}_v$, which is termed the *degree neighborhood of* v, is given by $\ddot{\mathbf{n}}_v = (\ddot{\mathrm{n}}_{k|v}, k \in \mathcal{D}_v)$, with entries $\ddot{\mathrm{n}}_{k|v} = \#\{u \in \Gamma_v | d_u = k\}$.

Proposition 1. *The entries* $C_{j,k}$ *of the conditional degree distribution matrix in Def. 4 are the*

averages of the fraction of neighbors of degree k over all degree j nodes:

$$C_{j,k} = \frac{1}{|\mathcal{V}_j|} \sum_{v \in \mathcal{V}_j} \frac{\ddot{\mathsf{n}}_{k|v}}{j}. \tag{2.5}$$

Proof. Observe $K_{j,k} = \sum_{v \in \mathcal{V}_j} \ddot{\mathsf{n}}_{k|v}$ by partitioning all edges with a degree j endpoint by the degree of the other endpoint, and $\sum_{k' \in \mathcal{D}} K_{j,k'} = j|\mathcal{V}_j|$ since the sum is the number of edges with a degree j endpoint. □

Row j of $\mathbf{C}$ is denoted $\mathbf{C}_j = (C_{j,k}, k \in \mathcal{D})$. The locations of the non-zero entries of $\mathbf{C}_j$ are denoted $\ddot{\mathcal{D}}_j = \bigcup_{v \in \mathcal{V}_j} \mathcal{D}_v$, i.e., $k \in \ddot{\mathcal{D}}_j$ means there exists some $uv \in \mathcal{E}$ with $d_u = j$ and $d_v = k$.

2.4.2 Biased walk degree transition probability distribution

Define the biased walk degree transition probability distribution as $\breve{\mathbf{p}}(\ddot{\mathbf{n}})$ and average biased walk degree transition matrix as $\bar{\mathbf{P}}$.

Definition 5. *The biased random walk degree transition probability distribution* $\breve{\mathbf{p}}(\ddot{\mathbf{n}}_v) = (\breve{p}_k(\ddot{\mathbf{n}}_v), k \in \mathcal{D})$ *from a node with degree neighborhood* $\ddot{\mathbf{n}}_v = (\ddot{\mathsf{n}}_{k|v}, k \in \mathcal{D})$ *(with* $\ddot{\mathsf{n}}_{k|v} = |\{u \in \Gamma_v | d_v = k\}|$*) is*

$$\breve{p}_k(\ddot{\mathbf{n}}_v) = \frac{\ddot{\mathsf{n}}_{k|v} l^\beta}{\sum_{k' \in \mathcal{D}} \ddot{\mathsf{n}}_{k'|v} k'^\beta}. \tag{2.6}$$

The following is an immediate result of Def. 2.

Proposition 2. *The biased random walk V in Def. 2 obeys the degree transition probability*

$$\mathbb{P}(d(V(t+1)) = k | V(t) = v) = \breve{p}_k(\ddot{\mathbf{n}}_v), \ k \in \mathcal{D}_v. \tag{2.7}$$

The above proposition states that the biased random walk V degree transition probability distribution has the property that the probability of the degree of the next node of the biased random walk depends upon the current node v only through $\ddot{\mathbf{n}}_v$. The next definition gives an average transition probability from nodes of degree j to nodes of degree k under the biased random walk.

Definition 6. *The average biased walk degree transition matrix* $\bar{\mathbf{P}}_V$ *is the* $\omega \times \omega$ *matrix with entries*

$$\bar{\mathbf{P}}_V(j,k) = \frac{1}{|\mathcal{V}_j|} \sum_{v \in \mathcal{V}_j} \breve{p}_k(\ddot{\mathbf{n}}_v), \ (j,k) \in \mathcal{D}^2 \tag{2.8}$$

giving the average probability of transitioning to a node of degree k over all starting nodes of degree j.

A key point, developed below, is that $\bar{\mathbf{P}}_V$ is of size $\omega \times \omega$ whereas the transition matrix $\mathbf{P}_V$ is of (potentially) significantly larger size $n \times n$.

2.4.3 Approximate biased walk degree transition matrix

A random ω-vector $\mathbf{N} = (\mathsf{N}_k, k \in \mathcal{D})$ has a *Multinomial distribution* with parameters $(j, \mathbf{p})$, denoted $\mathbf{N} \sim \text{Mult}(j, \mathbf{p})$, if

$$\mathbb{P}(\mathbf{N} = \ddot{\mathbf{n}}) = \binom{j}{\ddot{\mathbf{n}}} \prod_{k \in \mathcal{D}} p_k^{\ddot{\mathsf{n}}_k}, \ \ddot{\mathbf{n}} \in \mathcal{N}_j, \tag{2.9}$$

where $\binom{j}{\ddot{\mathbf{n}}} = \binom{j}{\prod_{k \in \mathcal{D}} \ddot{\mathsf{n}}_k!}$ is the Multinomial coefficient. Here, $\mathbf{N}$ has support $\mathcal{N}_j = \{\ddot{\mathbf{n}} \in \mathbb{N}^\omega | \ddot{\mathbf{n}} = (\ddot{\mathsf{n}}_k, k \in \mathcal{D}), \sum_{k \in \mathcal{D}} \ddot{\mathsf{n}}_k = j\}$, defined as the set of all possible ω-vectors from $\mathbb{N}^\omega$ that sum to j. Notice that vectors $\ddot{\mathbf{n}}$ which make up the support of $\mathbf{N}$ are arbitrary in that they are independent of a particular graph G or node v.

The approximate biased walk degree transition matrix $\tilde{\mathbf{P}}_Z$ has entries defined using the expectation of the biased random walks degree transition probability distribution $\breve{\mathbf{p}}_l$ when the degree neighborhood $\ddot{\mathbf{n}}$ is taken as a random Multinomial vector $\mathbf{N}$ with parameters $(j, \mathbf{C}_j)$.

Definition 7. *The approximate biased walk degree transition matrix $\tilde{\mathbf{P}}_Z$ is the $\omega \times \omega$ matrix with entries*

$$\tilde{\mathbf{P}}_Z(j,k) = \mathbb{E}\left[\breve{p}_k(\mathbf{N})\right], \ \mathbf{N} \sim \text{mult}(j, \mathbf{C}_j). \tag{2.10}$$

That is:

$$\tilde{\mathbf{P}}_Z(j,k) = \sum_{\ddot{\mathbf{n}} \in \mathcal{N}_j} \binom{j}{\ddot{\mathbf{n}}} \prod_{k \in \breve{\mathcal{D}}_j} (C_{j,k})^{\ddot{\mathsf{n}}_k} \frac{\ddot{\mathsf{n}}_k k^\beta}{\sum_{k' \in \mathcal{D}_k} \ddot{\mathsf{n}}_{k'} k'^\beta}. \tag{2.11}$$

The model of a BRW given by $\tilde{\mathbf{P}}_Z(j,k)$ is suitable for those graphs for which $\bar{\mathbf{P}}_V(j,k) \approx \tilde{\mathbf{P}}_Z(j,k)$ for each $(j,k) \in \mathcal{D}^2$; as will be shown in Sec. 2.5, it is not difficult to identify graphs for which the model works, and to find graphs for which it does not. Finally, a biased random walk is defined on degree set $\mathcal{D}$.

Definition 8. *The absorbing discrete-time Markov chain (DTMC) $Z = (Z(t), t \in \mathbb{N})$ on $\mathcal{D}$ with bias parameter $\beta \geq 0$ has states $\mathcal{D}$, partitioned into absorbing states $\hat{\mathcal{D}} = \{\lambda\}$ and transient states $\breve{\mathcal{D}} = \mathcal{D} \backslash \hat{\mathcal{D}}$. The $\omega \times \omega$ transition probability matrix $\tilde{\mathbf{P}}_Z$ is given in Def. 7. The absorption times from each initial transient state are $\mathscr{T}_Z = (\mathsf{T}_Z(k), k \in \breve{\mathcal{D}})$, with $\mathsf{T}_Z(k) = \min\{t \in \mathbb{N}, Z(t) = \lambda | Z(0) = k\}$.*

Recall that for large graphs (large n) it is not possible to compute $\mathbb{E}[\mathsf{T}_V]$ and $\text{Var}(\mathsf{T}_V)$ for T_V in Def. 2 using Thm. 1 since the fundamental matrix $\mathbf{N}_V = (\mathbf{I}_{\breve{n}} - \mathbf{L}_{\breve{n}})^{-1}$ cannot be computed, on account of the difficulty of inverting the (large) $\breve{n} \times \breve{n}$ matrix $\mathbf{I}_{\breve{n}} - \mathbf{L}_{\breve{n}}$. In contrast, for graphs with

bounded ω, it *is* possible to compute $\mathbb{E}[\mathsf{T}_Z]$ and $\mathrm{Var}(\mathsf{T}_Z)$ for T_Z in Def. 8 using Thm. 1 since the corresponding fundamental matrix $\mathbf{N}_Z = (\mathbf{I}_{\omega-1} - \mathbf{L}_{\omega-1})^{-1}$ is obtained by inverting the (smaller) $(\omega-1)\times(\omega-1)$ matrix $\mathbf{I}_{\omega-1} - \mathbf{L}_{\omega-1}$.

The model's approximation $\mathsf{T}_V \approx \mathsf{T}_Z$ lies with the required approximation $\bar{\mathbf{P}}_V \approx \tilde{\mathbf{P}}_Z$ mentioned above. That is, *i*) although the biased random walk on the graph has a probability of transitioning to nodes of each degree $k \in \mathcal{D}_v$ from a node v of degree j that depends upon the exact degree neighborhood $\ddot{\mathbf{n}}_v$ (i.e., $\breve{p}_l(\ddot{\mathbf{n}}_v)$), *ii*) this probability is apporximated using the expectation with respect to a *random* degree neighborhood $\mathbf{N}$, drawn with parameters j and $\mathbf{C}_j$. This distribution $\mathbf{C}_j$ is the average distribution of the number of nodes of each degree that are neighbors of a randomly selected degree j nodes (Prop. 1).

2.5 Results

The experimentation framework for this chapter is written using the igraph Python library Csardi and Nepusz [2006]. The experimentation framework to generates instances of a graph family, then for each graph the random times required to find a maximum degree node is measured for *i*) a biased random walk (BRW), and using a *ii*) random sampling. Unless noted otherwise, the mean $\mathbb{E}[\mathsf{T}_V]$ and standard deviation $Std[\mathsf{T}_V]$ of the absorption of a BRW on a graph is calculated from 500 trials for each value of the bias coefficient β. These trials compare the empirical mean $\mathbb{E}[\mathsf{T}_V]$ and empirical standard deviation $Std[\mathsf{T}_V]$ measured from the BRW on the graph with the numerical mean $\mathbb{E}[\mathsf{T}_Z]$ and standard deviation $Std[\mathsf{T}_Z]$ of a BRW on the Markov chain of the graph's degree states. Tab. 2.1 gives statistics for the Erdős-Rényi (ER) graphs the trials where conducted on.

Table 2.1: Parameters of the graphs used in the simulations.

Graph	Parameters	Size	λ
Erdős-Rényi (ER)	$p = 0.05$	100	11
Erdős-Rényi (ER)	$p = 0.0024$	1121	10
Erdős-Rényi (ER)	$p = 0.002569$	1011	11

As implemented the model has a significant computational limitation in that the approximate random walk ($Z(t)$) in Def. 8 requires computing the matrix $\tilde{\mathbf{P}}_Z$ in Eq. (2.11), and each such entry $\tilde{\mathbf{P}}_Z(j,k)$ requires summing over all $\ddot{\mathbf{n}} \in \mathcal{N}_j$. As the size of $|\mathcal{N}_j|$ grows exponentially in j, the experimentation framework is unable to compute $\mathcal{N}_j$ for graphs with $\lambda > 15$; this is the motivation behind selecting (n,s) pairs for the ER graphs so that λ is small.

2.5.1 Erdős-Rényi (ER) Graphs

The Erdős and Rényi [1959] graph $G_\epsilon(n, s)$ is a family of random graphs with n nodes and each of the $\binom{n}{2}$ possible edges is added independently with probability s. The resulting graph has a binomial degree distribution, $\mathsf{w} \sim \text{Bin}(n-1, s)$. Running trials on larger graphs required ensuring the graph is constructed so as to have a bounded expected maximum degree, on account of the computational limitation $\lambda_\epsilon \leq 15$ discussed above. The following proposition gives an upper bound on the expected maximum of n IID random variables:

Proposition 3. *Let* $(\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$ *be IID with moment generating function (MGF)* $\phi(t) = \mathbb{E}[\mathrm{e}^{t\mathsf{Y}}]$ *and let* $\mathsf{Y}_{\max} = \max(\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$. *Then* $\mathbb{E}[\mathsf{Y}_{\max}] \leq \frac{1}{t}\log(n\phi(t))$.

It may be proved by application of Jensen's inequality to establish $\mathrm{e}^{t\mathbb{E}[\mathsf{Y}_{\max}]} \leq n\phi(t)$.

Applying the above rule to the random degrees $\mathbf{d} = (\mathsf{d}_v, v \in [n])$ of an ER graph $G(n, s)$, each $\mathsf{d}_v \sim \text{Bin}(n-1, s)$, which are identically distributed, but not independent; it can be shown that the slight dependence is inessential and the bound applies. Recall that the binomial distribution $\text{Bin}(n, c/n)$ can be approximated as a Poisson distribution $\text{Poi}(c)$ in the case when n is large, and that the Poisson MGF is $\phi(t) = \mathrm{e}^{c(\mathrm{e}^t - 1)}$. Applying Prop. 3 to this case yields the following upper bound on the max degree of an ER graph $G(n, c/n)$

$$\mathbb{E}[\boldsymbol{\lambda}_\epsilon] \leq \frac{\log(n) - c}{W\left(\frac{\log(n)-c}{\mathrm{e}c}\right)} \tag{2.12}$$

where $W(\cdot)$ is the Lambert W function and $\boldsymbol{\lambda}_\epsilon$ is the maximum degree of a random ER graph. The value of Eq. (2.12) is that it allows $c(n)$ to be selected so that the resulting ER graph of order n has a specified expected max degree upper bound.

However, the ER graph is known to be disconnected with high probability when $s(n) = c/n$ (or smaller) for any c, and a random walk is only guaranteed to find a maximum degree node for a connected graph. It is further known that an ER graph with $s(n) = c(\beta)/n$ will have a fraction β of the n nodes in the giant connected component where $c(\beta) = -\log(1-\beta)/\beta$ Janson and Luczak [2000]. Hence the trade-off faced in generating ER graphs is between a large fraction β in the giant connected component vs. a bounded max degree. For $n = 1090$ and $c = 2.8$ (i.e., $s(n) = c/n = 0.0024569$), the expected max degree upper bound in Eq. (2.12) is $\lambda_\epsilon \leq 11.1041$, and $\beta = 0.924975$ obeys $-\log(1-\beta)/\beta = c$, meaning the giant connected component will contain approximately $n\beta = 1008$ nodes.

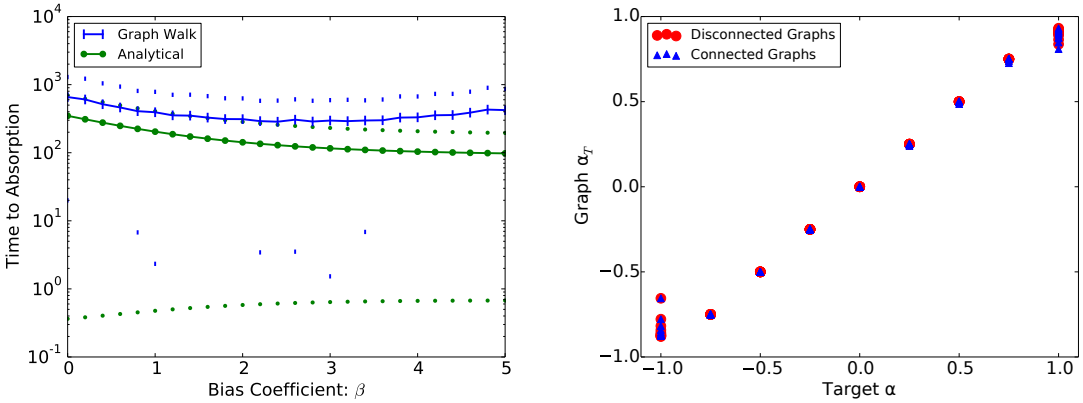

Figure 2.2: Left: average absorbtion time, $\mathbb{E}[\mathsf{T}]$ (solid lines), for original graph (blue, via simulation) and model (green, via (2.3)), with $\mathbb{E}[\mathsf{T}] \pm \mathrm{Std}[\mathsf{T}]$ (dashed lines). **Right:** output assortativity α_T as a function of input target assortativity α in the random rewiring algorithm.

2.5.2 Results for ER Graphs

The expected absorption time $\mathbb{E}[\mathsf{T}]$ for an ER graph using both *i*) the simulated biased random walk (BRW) on the original graph $\mathbb{E}[\mathsf{T}_V]$, and *ii*) the analytically computed $\mathbb{E}[\mathsf{T}_Z]$ using the reduced state space model, both swept over a range of bias coefficients β are shown in Fig. 2.2 (left). The plot shows a significant deviation between the measured quantity and the model prediction. The failure of the model for this graph may be accounted for by the fact that the ER graph is known to have zero assortativity, i.e., the degrees of the two endpoints of the graph are conditionally independent, and as such the degree of the current node does not provide any substantial information about the proximity of that node to higher degree (and by extension, maximum degree) nodes. In this sense, the fact that the DTMC model Z breaks down for such graphs, reflects the central assumptions of the model, which is that the degree of a node *does* contain information about its degree neighborhood and all nodes of degree k have similar degree neighborhoods.

This leads to the set of questions introduced at the start of this chapter: *i*) is the accuracy of the DTMC model W of a BRW on a graph dependent upon the graph's assortativity?, *ii*) are there graphs where BRW finds the max degree nodes faster than random sampling?, and *iii*) what is the optimal bias coefficient β^* for a graph of assortativity α?

2.5.3 Graph Re-wiring Algorithm

The following sections offer some preliminary numerical answers to these questions, using graph assortativity, denoted by $\alpha \in [-1, +1]$, as the independent control parameter. In these sections a

modified version of Brunet's rewiring algorithm Xulvi-Brunet and Sokolov [2005] for increasing or decreasing a graph's assortativity was used to construct graphs with a target assortativity, α_T.

The algorithm operates as follows. Given an initial graph G_0 and a target assortativity α_T. It calculates α_0, G_0's assortativity, then chooses two edges at random e_1, e_2 which are removed from G_0. If two new edges e_3 and e_4 can be wired between the endpoints of the former edges e_1 and e_2 without creating self loops or multiple edges and the assortativity of the new graph α_1 is closer to α_T than α_0, edges e_3 and e_4 are added, otherwise edges e_1 and e_2 are replaced. This procedure is repeated until the graph's assortativity is within a suitably small interval around α_T. Notice that this procedure preserves the degree distribution of G_0, since the degree of the end points of e_1 and e_2 are unchanged. When the assortativity converges, if the graph is disconnected, then for each disconnected component a random node is selected in the graph's giant component and wired to the smaller component, thereby connecting the graph. The assortativity of the 1011 node graphs compared to their target assortativity is shown in Fig. 2.2 (right) for both disconnected (red) and connected (blue) graphs. From Fig. 2.2 (right) it can be inferred that connecting sufficiently dense graphs in this manner has little effect on their assortativity. Similarly it can be shown numerically that for large n the effect of heuristically connecting disconnected graphs has limited impact on the graphs degree distribution.

2.5.4 Biased Random Walks

The following results are for Monte carlo simulations consisting of 500 trials on 10 graphs with binomial degree distributions of 100 and 1011 nodes, while sweeping the bias coefficient β of the walk. These simulations compared BRWs with two random sampling algorithms: *i*) sampling nodes without replacement, denoted 'RSW', and *ii*) sampling a node and all of its neighbors without replacing the sampled node, denoted 'SS-C'. The rationale for these two forms of sampling is in the interest of fairly comparison the absorption time between the BRW and a random sample. The BRW algorithm presumes at each step that the search is able to not only view the degree of the current node but also the degree of all neighbors of that node. Thus any comparison between the performance of, say, k steps of the BRW and k nodes sampled without replacement is unfair to random sampling, since the latter does not see as many nodes as the former. The second sampling scheme, SS-C, corrects this issue since in each step of SS-C a center node is selected at random, its degree as well as the degrees of its neighbors are sampled, before the center node is removed from the sample set; meaning both SS-C and BRW sample a similar number of nodes per sampling step.

The simulations compared BRW and random sampling RSW and SS-C graphs with target as-

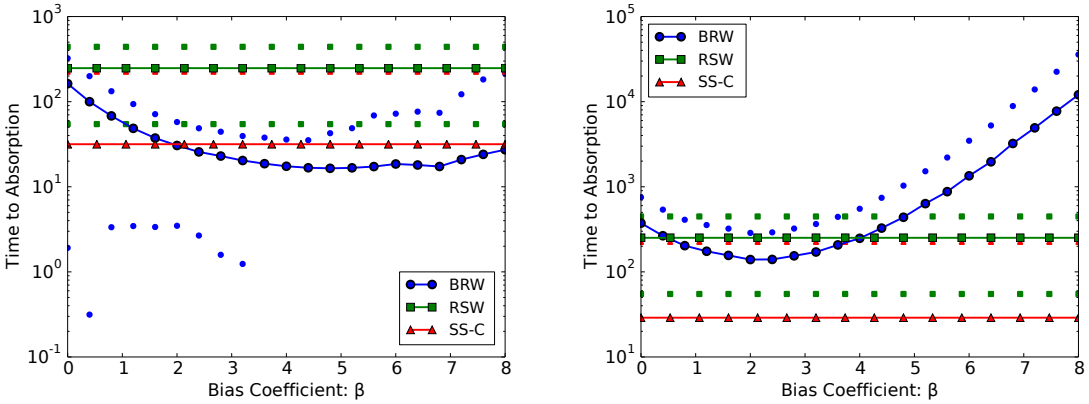

Figure 2.3: ER graph with ~ 1000 nodes. Expected time to absorption $\mathbb{E}[\mathsf{T}]$ (solid) and $\mathbb{E}[\mathsf{T}] \pm Std(\mathsf{T})$ (dotted) for i) the BRW (blue) and ii) random sampling without replacement (observing a) just the degree of the sampled node (green) and b) degrees of node and its neighbors (red)) versus β. $\alpha_T = +0.5$ (left) and $\alpha_T = -0.5$ (right).

sortativity $\alpha_T \in \{-1.0, -0.75, \ldots, 0.75, 1.0\}$ for both 100 node and 1000 node ER graphs. For each target α_T, and each graph rewired to that α_T, the bias coefficient β was swept over the range $[0, 8]$.

The results for $\alpha_T \in \{+0.5, -0.5\}$ are shown in Fig. 2.3 (for $n = 1000$) and Fig. 2.4 (for $n = 100$). Several points bear mention. First, for target assortativity $\alpha_T = +0.5$ (both for $n \approx 1000$ and $n \approx 100$) there exist optimized $\beta^*(\alpha_T)$ for which BRW outperforms random sampling of a node and its neighbor degrees, and there exist non-optimal values of β for which random sampling outperforms the non-optimized BRW. Second, for target assortativity $\alpha_T = -0.5$, the BRW is inferior to random sampling a node and its neighbor degrees for *all* values of β. Third, the case of $n \approx 1000$ and $\alpha = +0.5$ shows there exists a non-trivial value of β^* for this α, meaning the optimal value is not at either endpoint of the β interval of $[0, 8]$. This suggests that although BRWs can yield superior search times compared with sampling neighborhoods, doing so requires a correctly-tuned value of β for the particular value of α (and in this case also n). The inferiority of the BRW for disassortative graphs (here, $\alpha = -0.5$) may be on account of the fact that BRWs on such graphs are more likely to spend much of their search time trapped in a local minimum.

2.5.5 Optimal Bias Coefficient β

In this subsection the target assortativity α_T is extended to include the nine values mentioned earlier. For each α_T the optimal bias, $\beta^*(\alpha_T)$, is determined which gives $\mathbb{E}[\mathsf{T}^*] \equiv \mathbb{E}[\mathsf{T}(\alpha_T, \beta^*)]$, the expected time to absorption for the graph given the optimal bias coefficient. Both of these functions are plotted against α_T. Because the dependence of $\mathbb{E}[\mathsf{T}]$ on β can be somewhat flat near

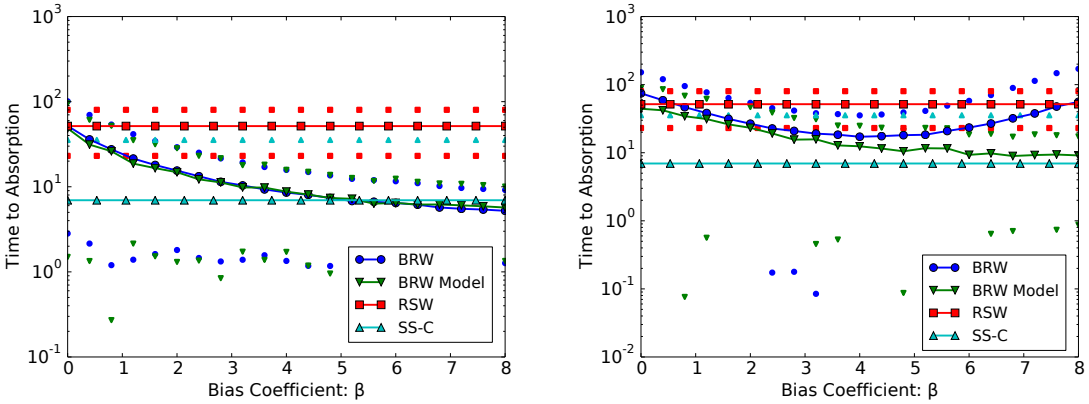

Figure 2.4: Same caption as Fig. 2.3 but for an ER graph with ~ 100 nodes. $\alpha_T = +0.5$ (left) and $\alpha_T = -0.5$ (right). Also shown is mean absorption time $\mathbb{E}[\mathsf{T}_Z]$ predicted by the model.

the optimal, meaning there is some degree of insensitivity to the precise value of β, therefore the interval $[\beta_-(\alpha_T), \beta_+(\alpha_T)]$ is computed containing $\beta^*(\alpha_T)$, where the interval holds all values of β for which the corresponding value of $\mathbb{E}[\mathsf{T}]$ is within 10% of the optimal value $\mathbb{E}[\mathsf{T}^*]$.

The results are shown in Fig. 2.5 and Fig. 2.6. Several points bear mention. First, Fig. 2.5 shows that the optimal bias coefficient β^* for $\beta \in [0, \ldots, 8]$ tends to increase with increasing assortativity α_T of the graph. More sample graphs for each α_T, more points $\alpha_T \in [-1, +1]$, and a larger search range for β than the current $[0, 8]$ are required to confirm this initial observation. Second, Fig. 2.6 shows that for both $n \approx 100$ and $n \approx 1000$ there exists an interval of α over which optimized BRWs outperform random sampling of node and neighbor degrees. Again, more extensive simulations are required. However, these preliminary results suggest BRWs are inferior to sampling for graphs with negative assortativity.

2.5.6 Biased Random Walk Model

Finally, the two right side plots in Fig. 2.5 and Fig. 2.6 include results for both the simulations of BRWs on the graph as well as analytical computations using the reduced state space model. In particular, Fig. 2.5 shows that β^* for minimizing $\mathbb{E}[\mathsf{T}_V]$ and β^* for minimizing $\mathbb{E}[\mathsf{T}_Z]$ are not exactly equal, but are comparable, and show the same rough increasing trend as a function of α. Moreover, Fig. 2.6 shows $\mathbb{E}[\mathsf{T}_V^*]$ is comparable to $\mathbb{E}[\mathsf{T}_Z^*]$ for certain values of α.

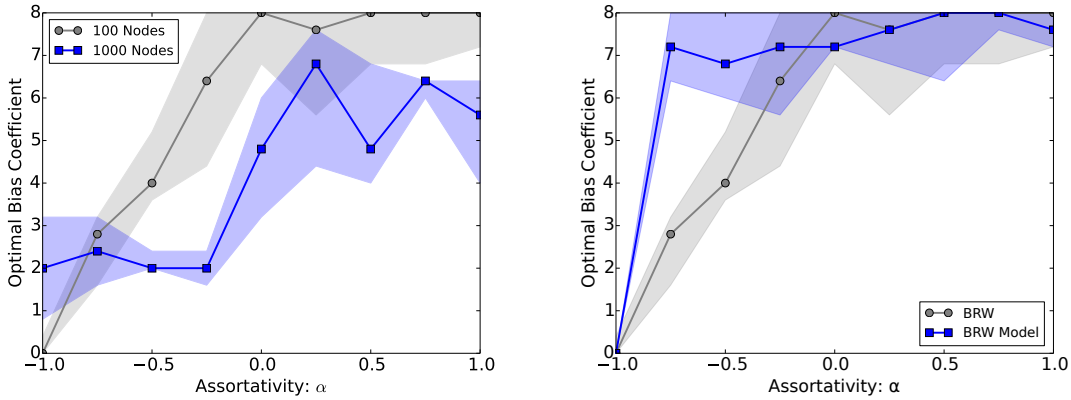

Figure 2.5: The optimal bias coefficient β^* (points) and the interval $[\beta_-, \beta_+]$ of points for which $\mathbb{E}[\mathsf{T}]$ is within 10% of $\mathbb{E}[\mathsf{T}^*]$ (shaded) vs. the target assortativity α_T. **Left:** comparison of $n \approx 100$ (gray) and $n \approx 1000$ (blue). **Right:** comparison of β^* for $n = 100$ for actual graph (gray) and reduced state space BRW Model (blue).

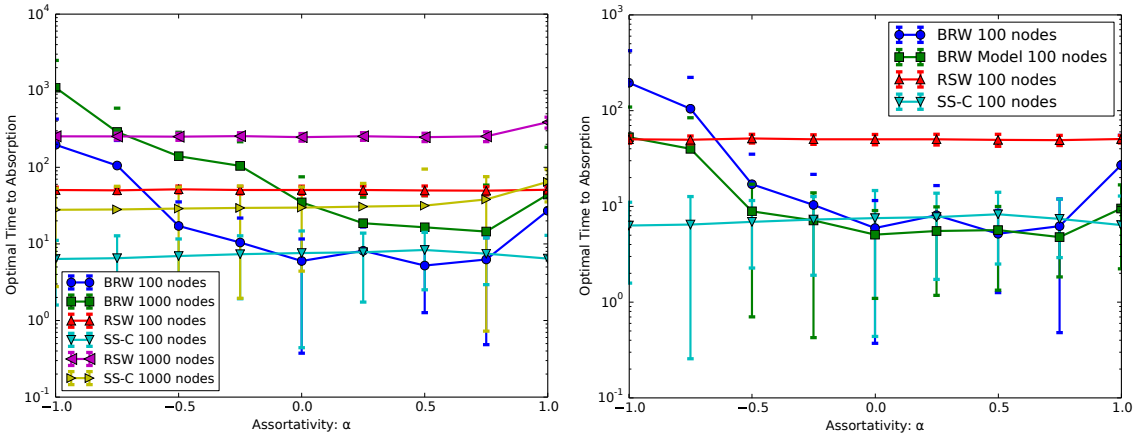

Figure 2.6: Mean absorption times $\mathbb{E}[\mathsf{T}^*]$ vs. α_T, using the optimized value $\beta^*(\alpha_T)$ from Fig. 2.5. **Left:** comparison of $n \approx 100$ (blue) and $n \approx 1000$ (green). **Right:** comparison of $\mathbb{E}[\mathsf{T}^*]$ for actual graph (blue) and reduced state space BRW Model (green).

Chapter 3: On random walks and random sampling to find max degree node in assortative Erdős Rényi graphs

3.1 Introduction

3.1.1 Motivation

The increasing prevalence and size of social networks motivates interest in efficiently searching these networks for information. It is natural to model a social network as a graph, with nodes representing users and edges representing associations between users. One can define the "popularity" or "influence" of a user via the degree of its corresponding node in the graph. A natural objective is to find one of the most popular users, i.e., to identify a max degree node. As these networks are large, this requires an efficient search algorithm designed to minimize the time required to find such a node. The identity and location of max degree nodes has applications in the study of percolation on complex networks, instances of which include the spread of epidemics through populations and the dispersion of influence on social networks.

This chapter addresses two key questions via the widely used framework of Erdős Rényi (ER) random graphs. First, given that one is interested in finding a max degree node in a graph, it is prudent to study the question of how many such nodes there are, on average. Second, this chapter compares the search time of two basic random search algorithms / paradigms: random walks vs. random sampling. A biased random walk on a graph involves repeatedly selecting a neighboring node of the current node until a target is reached, while random sampling involves repeatedly selecting a node (with or without replacement) until a target is reached. These two approaches differ in that the former exploits the local edge structure of the graph while the latter ignores it. The second question addressed can thus rephrased as asking whether local knowledge of a graphs structure may be used to reduce the search time relative to sampling.

3.1.2 Prior and Related Work

There are two main sections of this chapter, Sec. 3.2 which looks at the expected number of maximizers in a set of binomial RV's, and Sec. 3.3 which examines how the time a biased random walk takes to find a max degree node is influenced by the assortativity of the underlying graph. The expected number of maximizers in a vector of n IID discrete[1] RVs has been addressed in the literature for

[1]The maximizer of a vector of IID. continuous RVs is almost surely unique.

over twenty years; this is only the most cursory summary of this literature due to space constraints. Early papers include Eisenberg et al. [1993], Brands et al. [1994], Baryshnikov et al. [1995], Qi and Wilms [1997], Qi [1997], and Olofsson [1999], while more recent papers includes Bruss and Grübel [2003] and Eisenberg [2008, 2009]. A subtle but important point about most of this literature is that many of the above works assume an *unbounded* support for the RVs, with the justification being that finite support distributions have degenerate behavior of K_n, namely $\mathsf{K}_n \to \infty$ almost surely. This is a false dichotomy, however, as it omits the case where the distribution is finite but growing in n, as with the binomial distribution, with the consequence that some existing results to do not apply to the binomial case. Additional portions of existing work are likewise not applicable as they address related but distinct problems, e.g., the asymptotic probability of a tie, i.e., $\lim_{n\to\infty} \mathbb{P}(\mathsf{K}_n > 1)$, and a focus on the case of geometric RVs. The focus in this chapter on $\mathbb{E}[\mathsf{K}(n)]$ for n IID. binomial RVs, each with support $\{0, \ldots, n-1\}$, is, at the time of writing, not yet addressed in the literature.

There is a growing body of work on the problem of finding a maximum degree node in a graph. Ikeda and Kubo [2003] proved a lower bound of $\Omega(n^2)$ for any algorithm to find a particular node on a graph of order n, and established that a biased random walk achieves $O(n^2)$. More recently, Cooper et al. Cooper et al. [2012] have shown that all nodes in a power law graph of degree n^a or higher can be found in $O(n^{1-(4/3)a+\delta})$ steps using biased random walks. Recently Avrachenkov introduced two algorithms for finding maximum degree nodes. The first is a random walk (RW) with uniform restart probability Avrachenkov et al. [2012], the second is a novel random sampling algorithm Avrachenkov et al. [2014]. The key distinctions between these algorithms and the degree biased random walks (BRW) studied in Sec. 3.3.2 is neither of Avrachenkov's algorithms sample the degree neighborhood of a node. BRW's however, must know the degree neighborhood of a sampled node to calculate the transition probabilities for the next step in the random walk. This may seem a trivial difference, but in fact it is a fundamental distinction between the sampling cost models assumed by Avrachenkov and a BRW. BRW's implicitly assume that the cost of sampling a node neighborhood is invariant in neighborhood size. Avrachenkov's RW makes no such assumption as it does not consider the neighborhood of a node and his random sampling scheme explicitly assumes the cost of sampling a node neighborhood scales in the size of that neighborhood.

The work presented in this chapter is distinct in that *i*) the figure of merit is to find *any* maximum degree node, as opposed to one particular or all such nodes, and *ii*) the framework is probabilistic average case analysis (as opposed to worst-case algorithm complexity), using the ER family of random graphs. The results presented in this chapter build on those of Chap. 2 and suggest that the ability of a BRW to find a max degree node in fewer steps than random sampling is correlated

with a graph's assortativity.

3.1.3 Contributions and Outline

The most important results presented in this chapter are as follows. First, Sec. 3.2 proves that the average number of maximum degree nodes in a large ER graph is one, which exacerbates the difficulty of the search problem to one analogous to "finding a needle in a haystack". Second, Sec. 3.3 investigates the search time of both random walk and random sampling in as a function of the assortativity of the graph, and show a connection between the search time of a random walk and the prevalence of locally maximum degree nodes. A brief conclusion is given in Sec. 3.4 and some additional proofs are in Sec. 3.5.

3.2 Expected number of max degree nodes in ER graphs

3.2.1 Preliminaries

Let $G = (\mathcal{V}, \mathcal{E})$ denote an undirected graph with $\mathcal{V} = [n]$ and edge set $\mathcal{E}$. Let $\Gamma_i = \{j | ij \in \mathcal{E}\} \subseteq [n] \setminus i$ be the neighbors and $d_i = |\Gamma_i|$ the degree of node $i \in [n]$. Define *i)* $\mathcal{D} = \bigcup_{i \in [n]} d_i$ as the set of degrees found in G, *ii)* $\mathbf{n} = (n_k, k \in \mathcal{D})$ as the number of nodes of each degree, *iii)* $\mathbf{w} = (w_k, k \in \mathcal{D})$ with $w_k = n_k/n$ as the probability a randomly selected node has degree k, and *iv)* $\mu = \sum_{k \in \mathcal{D}} k w_k$ the average degree.

An ER graph Erdős and Rényi [1959] with parameters (n, s) is a *random* undirected graph $G_\epsilon = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = [n]$, and random edge set $\mathcal{E}$, with each of the $\binom{n}{2}$ possible edges added independently with probability $s \in (0, 1)$. Let $\hat{\mathsf{X}}_i = d_i$ denote the random degree of node $i \in [n]$, and let $\hat{\mathbf{X}} = (\hat{\mathsf{X}}_i, i \in [n])$ be the random n-vector holding the degrees of each node. Then *i)* the maximum degree is the random variable (RV) $\hat{\mathsf{M}}(n) = \max(\hat{\mathbf{X}})$ (note $\hat{\mathsf{M}}(n) \in \{0, \ldots, n-1\}$), *ii)* the set of maximum degree nodes is the random set $\hat{\mathcal{K}}(n) = \text{argmax}(\hat{\mathbf{X}})$ (note $\hat{\mathcal{K}}(n) \subseteq [n]$), and *iii)* the number of maximum degree nodes is the RV $\hat{\mathsf{K}}(n) = |\hat{\mathcal{K}}(n)|$ (note $\hat{\mathsf{K}}(n) \in [n]$). Sec. 3.2 of this chapter studies $\mathbb{E}[\hat{\mathsf{K}}(n)]$ while Sec. 3.3 looks at the average time to find a node in $\hat{\mathcal{K}}(n)$ via BRW and via SS-S.

Given n nodes and an edge probability $s \in (0, 1)$, ER graph $G_\epsilon = (\mathcal{V}, \mathcal{E})$ has a binomial degree distribution with parameters $(n-1, s)$. The RVs $\hat{\mathbf{X}} = (\hat{\mathsf{X}}_1, \ldots, \hat{\mathsf{X}}_n)$, with $\hat{\mathsf{X}}$ the random degree of a given node, have identical distribution $\hat{\mathsf{X}} \sim \text{Bin}(n-1, s)$, but are *not* independent: the *edges* are placed independently, with each placed edge incrementing the degrees of *both* of its endpoints. Although $\hat{\mathbf{X}}$ is not an independent vector, it is not hard to show that the positive correlation is inversely proportional to n, and that $\hat{\mathbf{X}}$ is *asymptotically* independent as $n \to \infty$. In this chapter

$\hat{\mathbf{X}}$ is approximated as $\mathbf{X}$, an independent vector with the same marginals $\hat{\mathsf{X}} \sim \mathsf{X}$, and show via simulation that the omitted dependence has a negligible impact on the quantities of interest. Given that the motivating objective of this thesis is to find one of the maximum degree nodes in a graph, it is natural to inquire as to the average *number* of such nodes. If one thinks of the nodes of a graph as the proverbial "haystack" and the max degree nodes as the "needles" therein, then one of the key questions addressed in this chapter is how may needles are in the haystack on average. Formally, defining $\mathsf{K}(n) = |\operatorname{argmax}(\mathsf{X}_1, \ldots, \mathsf{X}_n)|$ as the number of degree maximizers; the interest in this chapter is $\mathbb{E}[\mathsf{K}(n)]$.

3.2.2 Model and Preliminary Results

Write $[n] \equiv \{0, \ldots, n-1\}$ for the support, i.e., $\mathsf{X} \in [n]$, denote the probability mass function (PMF) by $\mathbf{w} = (w_k, k \in [n])$ with entries $w_k = \binom{n-1}{k} s^k (1-s)^{n-1-k}$, and denote the cumulative distribution function (CDF) by $\mathbf{W} = (W_k, k \in [n])$, with entries $W_k = \sum_{j=0}^{k} w_j$. Define the PMF ratio $\mathbf{r} = (r_1, \ldots, r_{n-1})$ with entries $r_k = w_{k-1}/w_k > 0$, and the CDF ratio $\mathbf{R} = (R_1, \ldots, R_{n-1})$ with entries $R_k = W_k/W_{k-1} > 1$. Notice that $\mathbf{w}, \mathbf{W}, \mathbf{r}, \mathbf{R}$ have entries w_k, W_k, r_k, R_k that each depend upon n, although the notation does not make this explicit. The following lemma, which applies to an *arbitrary* discrete distribution, is well-known and easily proved (e.g., [Brands et al., 1994, Corollary 2.2]).

Lemma 1. *For a distribution with PMF* $\mathbf{w}$ *and CDF* $\mathbf{W}$*:*

$$\mathbb{E}[\mathsf{K}(n)] = n \sum_{k \in [n]} w_k W_k^{n-1}. \tag{3.1}$$

This chapters first result, Prop. 4, with proof in Sec. 3.5.2, applies to an arbitrary discrete distribution with increasing $\mathbf{r}$ and decreasing $\mathbf{R}$, and leverages Lem. 1 into two upper bounds on $\mathbb{E}[\mathsf{K}(n)]$.

Proposition 4. *If* $\mathbf{r}$ *is increasing in* k *and* $\mathbf{R}$ *is decreasing in* k *then, for all* $n \in \mathbb{N}$ *and* $k \in [n-1]$*,*

$$\begin{aligned} \mathbb{E}[\mathsf{K}(n)] &\leq u(k,n) \equiv r_k(W_k^n - w_0^n) + (1 - W_{k-1}^n) R_k^{n-1} \\ &\leq v(k,n) \equiv r_k + n(1 - W_{k-1}) R_k^n \end{aligned} \tag{3.2}$$

Lem. 2, the proof of which is in Sec. 3.5.2, establishes $\mathbf{r}, \mathbf{R}$ for the binomial distribution satisfy the requirements of Prop. 4.

Lemma 2. *For the binomial distribution* $\mathbf{r}$ *is increasing, with* $r_{\lfloor ns \rfloor} \leq 1 \leq r_{\lceil ns \rceil}$*, and* $\mathbf{R}$ *is decreasing.*

3.2.3 Main Results

Thm. 2, with proof in Sec. 3.5.2, is the main result of Sec. 3.2. It *i*) leverages Prop. 4 into an upper bound $w(k,n)$ which holds for the binomial case with $k = t(n-1)$ and n sufficiently large; *ii*) shows the convergence (in n) of the upper bound $w(k,n)$ from Prop. 4 for $k(n) = t(n-1)$, for $t \in (s,1)$; which in turn implies *iii*) $\lim_{n\to\infty} \mathbb{E}[\mathsf{K}_n] \leq 1+\epsilon$ for arbitrarily small $\epsilon > 0$. Thus, for sufficiently large ER graphs, *there is on average a unique maximum degree node*, for any choice of $s \in (0,1)$.

Theorem 2. *For the binomial distribution $\mathbb{E}[\mathsf{K}(n)] \leq y(t(n-1),n)$ for sufficiently large n, and $t \in (s,1)$, where*

$$y(k,n) \equiv r_k + nb(k,n)\mathrm{e}^{nb(k,n)}, \tag{3.3}$$

and $b(k,n) = \frac{r_{k+1}}{r_{k+1}-1} w_k$. Moreover, $y(t(n-1),n) \to y(t) \equiv \frac{t(1-s)}{(1-t)s}$ for $t \in (s,1)$, and in particular, $\lim_{n\to\infty} \mathbb{E}[\mathsf{K}(n)] \leq 1+\epsilon$, for arbitrary $\epsilon > 0$.

Remark 1. *There is a tension in the choice of $t \in (s,1)$ in Thm. 2 between the tightness of the asymptotic bound and the corresponding rate of convergence: the asymptotic bound $w(t)$ is decreasing in t as $t \downarrow s$, but the rate of convergence of $nb(t(n-1),n) \to 0$ is likewise decreasing in t as $t \downarrow s$.*

Remark 2. *After the class of ER graphs with fixed $s \in (0,1)$, perhaps the second most important class of ER graphs is those with $s(n) = c/(n-1)$ for $0 < c < n-1$. For n large the distribution $\mathsf{X} \sim \mathrm{Bin}(n-1, c/(n-1))$ approaches that of $\tilde{\mathsf{X}} \sim \tilde{\mathbf{p}} \equiv Poi(c)$, and thus $\mathsf{K}(n)$ is approximately $\tilde{\mathsf{K}}(n)$, the number of maxima of n IID Poisson RVs with parameter c. In this case the support of $\tilde{\mathsf{X}}$ is unbounded, and as such prior results from the literature apply. In particular, since $\tilde{w}_k/\tilde{w}_{k-1} \to 0$ as $k \to \infty$, Theorem 2.6 from Brands et al. [1994] guarantees $\tilde{\mathsf{K}}(n)$ does not converge in distribution.*

3.2.4 Numerical Results

Three plots illustrate the results in Sec. 3.2. Fig. 3.1 illustrates $\mathbb{E}[\mathsf{K}(n)]$ for $\mathbf{X}$ IID. with components $\mathsf{X} \sim \mathrm{Bin}(n-1,s)$, and $s \in (0,1)$, vs. $\mathbb{E}[\hat{\mathsf{K}}(n)]$ for $\hat{\mathbf{X}}$ the empirical degree distribution for an ER graph with the same parameters. The plot makes clear the slight dependence in $\hat{\mathbf{X}}$ has a negligible impact on $\mathbb{E}[\hat{\mathsf{K}}(n)]$, justifying the focus on the IID. case. Fig. 3.2 illustrates $\mathbb{E}[\mathsf{K}(n)]$ for the binomial case with constant s vs. $s \in (0,1)$, and the corresponding bounds $u(k,n), v(k,n)$ from Prop. 4, and the bound $y(k,n)$ from Thm. 2 (each with k numerically optimized for each (n,s) pair). All curves are for $n = 100$. Observe $\mathbb{E}[\mathsf{K}(n)] \in (1.29, 2)$ for any $s \in (0.02, 0.92)$ (shown as gridlines). Fig. 3.3 illustrates $\mathbb{E}[\mathsf{K}(n)]$ for the binomial case with $s(n) = c/(n-1)$ vs. $c \in [0.1, 10]$, along with the corresponding Poisson approximation $\mathbb{E}[\tilde{\mathsf{K}}(n)]$ (c.f. Rem. 2), and the optimized upper bound $u(k^*,n)$

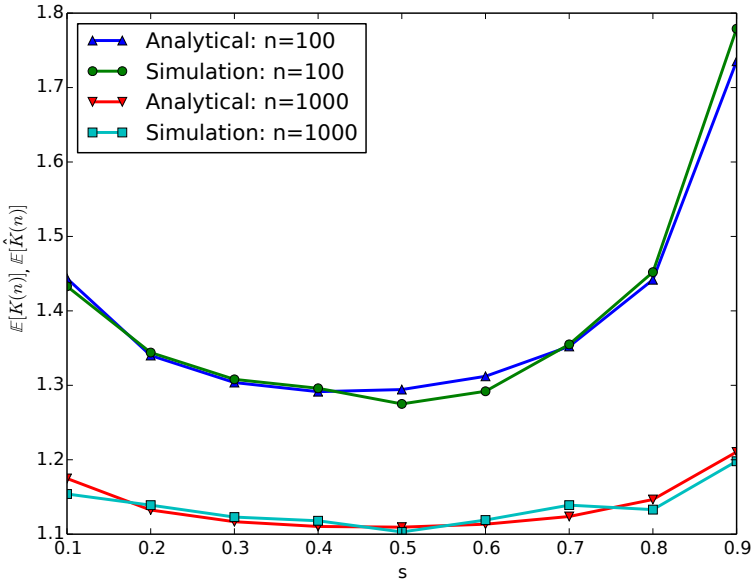

Figure 3.1: Compares the analytical quantity $\mathbb{E}[\mathsf{K}(n)]$ from Lem. 1 in Sec. 3.2 for the case of n IID to the average number over 1000 simulation of maximum degree nodes in an ER graph with parameters (n, s), denoted $\mathbb{E}[\hat{\mathsf{K}}(n)]$. binomial RVs with parameters $(n-1, s)$, versus the edge probability $s \in \{0.1, \ldots, 0.9\}$, for $n \in \{100, 1000\}$. The plot shows the degree dependence omitted in the analysis has a negligible impact, and the convergence of $\mathbb{E}[\mathsf{K}(n)]$ to 1 in n.

from Prop. 4. Observe that even though $\tilde{\mathsf{K}}(n)$ does not converge in distribution, the numerical results show $\mathbb{E}[\tilde{\mathsf{K}}(n)] < 2$ for any $c \geq 1$ and $n \leq 128$.

3.3 On random walk vs. random sampling in an assortative ER graph

3.3.1 Joint Degree Distribution and Assortativity

This section introduces the joint degree distribution $\mathbf{F} = \mathbf{F}(G)$ of an undirected graph $G = (\mathcal{V}, \mathcal{E})$, its relationship to the assortativity $\alpha = \alpha(G)$ of a graph, and a "rewiring" algorithm to transform an initial non-assortative graph into one with an approximate target assortativity $\alpha_T \in [-1, +1]$.

Let $\bar{\mathcal{D}} = \mathcal{D} \setminus \{0\}$ be the set of nonzero degrees in the graph, and define the $|\bar{\mathcal{D}}| \times |\bar{\mathcal{D}}|$ symmetric matrix $\mathbf{F}$ with entries

$$F_{jk} = \begin{cases} \frac{K_{j,k}}{2m}, & j \neq k \\ \frac{K_{j,j}}{m}, & j = k \end{cases}, \; j, k \in \bar{\mathcal{D}} \tag{3.4}$$

where $K_{j,k} = K_{k,j} = \#\{e \in \mathcal{E} | d(e) = \{j, k\}\}$[2], for $j, k \in \bar{\mathcal{D}}$ with $j \neq k$ the number of edges with endpoints with degrees j, k, and $K_{j,j} = \#\{uv \in \mathcal{E} | d_u = d_v = j\}$ the number of edges with both

[2] $d(e)$ denotes the unordered pair of degrees of the endpoints of edge $e \in \mathcal{E}$.

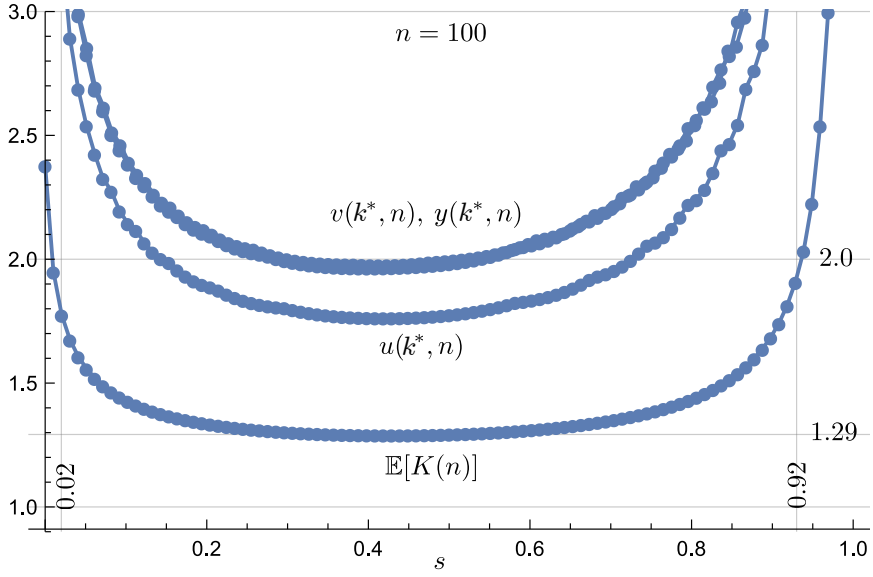

Figure 3.2: The quantity $\mathbb{E}[\mathsf{K}(n)]$ and the optimized bounds $u(k^*, n), v(k^*, n), y(k^*, n)$ from Prop. 4 and Thm. 2 in Sec. 3.2 for the binomial distribution with constant s vs. $s \in (0, 1)$. In ascending order, the curves are $\mathbb{E}[\mathsf{K}(n)]$, $u(k^*, n)$, $v(k^*, n)$, and $y(k^*, n)$, with the latter two visually indistinguishable.

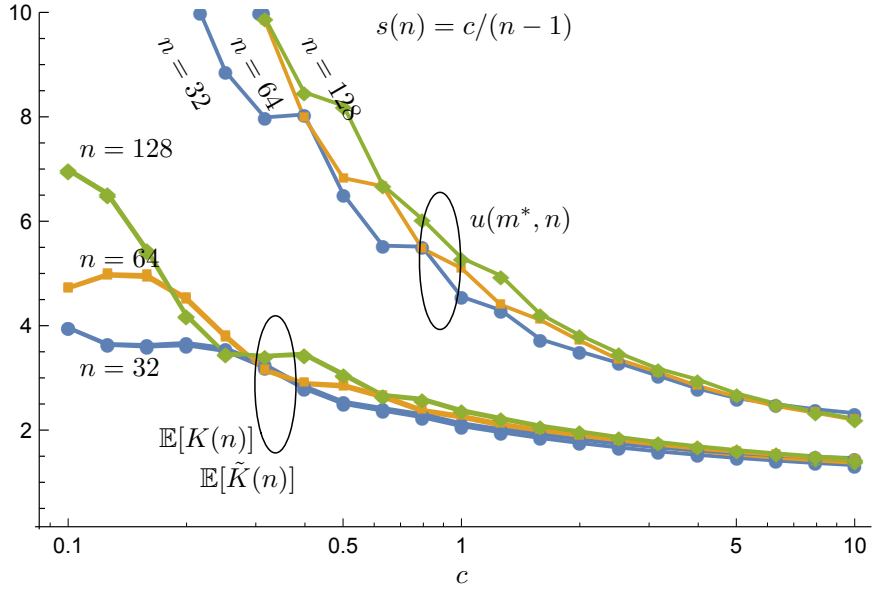

Figure 3.3: The quantity $\mathbb{E}[\mathsf{K}(n)]$ and its Poisson approximation $\mathbb{E}[\tilde{\mathsf{K}}(n)]$ (visually indistinguishable, bottom curves) along with the optimized upper bound $u(k^*, n)$ from Prop. 4 (top curves) for the binomial distribution with $s(n) = c/(n-1)$, vs. $c \in [0.1, 10]$, and $n \in \{32, 64, 128\}$ (blue, yellow, green), respectively.

endpoints of degree j. The matrix $\mathbf{F}$ is the *joint* degree distribution (thus $\sum_{j,k} F_{j,k} = 1$) of a randomly selected edge, and has row and column sums $\sum_k F_{j,k} = \sum_k F_{k,j} = f_j$, where $\mathbf{f} = (f_j, j \in \bar{\mathcal{D}})$ is a distribution with $f_j = jw_j/\mu$ representing the probability that a degree j node is chosen if a node is selected randomly from the two endpoints of a randomly chosen edge.

The graph assortativity Newman [2002] is the correlation of the endpoints of a randomly chosen edge and is given by

$$\alpha = \frac{1}{\sigma^2_{\mathbf{f}}} \sum_{j \in \mathcal{D}_0, k \in \mathcal{D}_0} jk(F_{j,k} - f_j f_k) \in [-1, +1], \tag{3.5}$$

where $\sigma^2_{\mathbf{f}} = \sum_{k \in \bar{\mathcal{D}}} k^2 f_k - \left(\sum_{k \in \bar{\mathcal{D}}} k f_k\right)^2$ is the variance of $\mathbf{f}$. Note that this definition differs from Newman [2002] original definition of a graphs degree assortativity in terms of the *excess* degree joint distribution, but it is not difficult to show that the two expressions are equivalent, see Sec. 3.5.1. Positive (negative) α indicates a tendency for the degrees of the endpoints of an edge to be similar (dissimilar), respectively.

It is easy to see that the assortativity $\alpha(G)$ of a random ER graph $G = (\mathcal{V}, \mathcal{E})$ has expected value zero for any n and converges in probability to zero as $n \to \infty$. ER graphs are non-assortative for the same reason behind the approximation of $\hat{\mathbf{X}}$ with $\mathbf{X}$ in Sec. 3.2: the degrees of the vertices are *nearly* independent, and asymptotically independent as $n \to \infty$.

As stated in the Sec. 3.1, a goal of this chapter is to compare BRW and SS-S‘ on ER graphs with *tunable* assortativity. This goal requires a means to *rewire* an ER graph to achieve (approximately) a target assortativity $\alpha_T \in [-1, +1]$, without disturbing the (marginal) degree distribution $\mathbf{w}$. The random rewiring algorithm (Alg. 1) is similar to others presented in the literature (e.g., Newman [2003], Xulvi-Brunet and Sokolov [2005]), and is included for completeness of presentation. The algorithm repeatedly selects a random pair of disjoint edges and rewires them in up to two different ways, as shown in Fig. 3.4, where a rewiring is *valid* provided the new edges are not already present in the graph. It then computes the associated assortativities and selects the best one, stopping when the rewired graph has assortativity within ϵ of the target α_T. The cumulative number of iterations required to reach various targets α_T (with the graph G_T for target α_T used as input for target α_{T+1}) is shown in Fig. 3.5 (left), and a comparison between the target and the measured assortativity is shown in Fig. 3.5 (right).

Because the objective of this section is to study the performance of a random walk, which requires a connected graph, the final stage of this algorithm connects any disconnected components, by adding edges at random edges between the giant component of G_T and each disconnected component. As shown in Fig. 3.5, this rewiring has a negligible impact on the assortativity. Finally, observe the

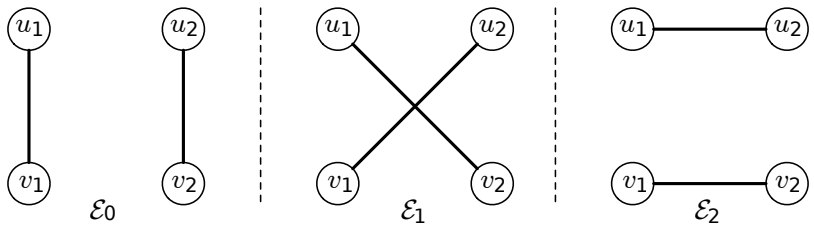

Figure 3.4: Illustration of the random rewiring in Alg. 1: two disjoint edges from E_0 are rewired in two different ways to form two new edge sets E_1, E_2. Each such rewiring is *valid* if it does not create multiple edges any pair of nodes.

rewiring algorithm, besides the final connecting of disconnected components, does not change the marginal degree distribution $\mathbf{w}$, but, of necessity, may significantly affect the *joint* degree distribution $\mathbf{F}$ in Eq. (3.4).

Algorithm 1 Rewire graph G to target assortativity α_T

---

require $G = (\mathcal{V}, \mathcal{E}), \alpha_T \in [-1, +1], \epsilon > 0$
initialize $G_0 = (\mathcal{V}, \mathcal{E}_0) \leftarrow G = (\mathcal{V}, \mathcal{E})$
while $|\alpha(G_0) - \alpha_T| > \epsilon$ **do**
    Pick random pair of disjoint edges $\{u_1, v_1\}, \{u_2, v_2\}$
    Rewire $\mathcal{E}_0$ into $\mathcal{E}_1$ and $\mathcal{E}_2$ as in Fig. 3.4
    Compute $\alpha_i = \alpha(G_i)$ for $i \in \{0, 1, 2\}$
    $\mathcal{E}_0 \leftarrow \mathcal{E}_{i^*}$ for $i^* = \text{argmin}_{i \in \{0,1,2\}} |\alpha_i - \alpha_t|$
end while
return $G_0 = (\mathcal{V}, \mathcal{E}_0)$

3.3.2 Biased Random Walk and Random Star Sampling

The rewiring Alg. 1 plays a key role in numerically investigating the average search time required of both biased random walks (BRW) and random star sampling without replacing the center node (SS-S) Kolaczyk [2009] to find a maximum degree node on graphs of varying assortativity. Formally, for any undirected and connected graph $G = (\mathcal{V}, \mathcal{E})$, and any randomized search algorithm (SA), define the RV, $\mathsf{T}(SA, G)$, as the number of steps required until SA discovers one of the maximum degree nodes in set $\mathcal{K}(G)$.

In random sampling without replacement (RSW), at each iteration a node is selected at random from the set of unsampled nodes and added to the sampled set. SS-S extends RSW so that at each step the selected node, say v, and all of its neighbors, Γ_v, are added to the sampled set. As will be made clear, SS-S offers a "fair" comparison with BRW.

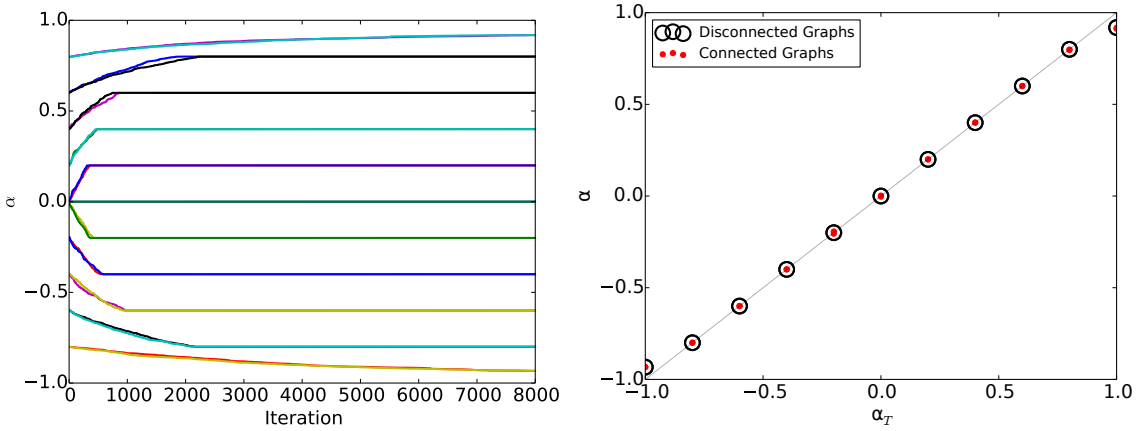

Figure 3.5: Left: α vs. iterations in Alg. 1 for targets $\alpha_T \in \{-1.0, -0.8, \ldots, 0.8, 1.0\}$. **Right:** target α_T and actual α from Alg. 1, before/after connecting components. In both plots $n = 500$ and $s = 0.01$.

The BRW is a RW on an undirected graph G with the transition probability

$$p_{uv} = \frac{d_v^\beta}{\sum_{v' \in \Gamma_u} d_{v'}^\beta}, \tag{3.6}$$

where $\beta \in [0, \infty]$ is the bias coefficient of the random walk Ikeda and Kubo [2003]. If $\beta = 0$ the BRW reduces to a uniform random walk, and as $\beta \to \infty$ it transitions with probability one to a randomly selected member of the set of highest degree nodes in Γ_u. The initial point of the walk v is selected at random among the set of nodes, $d_v = k$ with probability w_k. The BRW is guaranteed to reach the target set eventually on account of the assumption that the graph is connected. SS-S is the proper sampling analog of the BRW, in terms of nodes viewed per iteration, since computing the edge transition probability in a BRW requires the walk observe the degrees of each of the neighbors of the current node.

Let $\tilde{\mathsf{N}}$ denote the random search time for the BRW and $\bar{\mathsf{N}}$ the random search time of SS-S, and compare $\mathbb{E}[\tilde{\mathsf{N}}]$ and $\mathbb{E}[\bar{\mathsf{N}}]$ on ER graphs with assortativity tuned by Alg. 1 in Fig. 3.6. The key points are: *i*) the performance of SS-S, i.e., $\mathbb{E}[\bar{\mathsf{N}}]$, is independent of α, as is to be expected from the fact that Alg. 1 doesn't affect the marginal degree distribution $\mathbf{w}$; *ii*) the performance of BRW, i.e., $\mathbb{E}[\tilde{N}]$, is decreasing in α; *iii*) BRW is superior to SS-S for $\alpha \geq 0.34$. Recalling the question from Sec. 3.1, one may leverage the edges of the graph to improve the search time for graphs with positive assortativity.

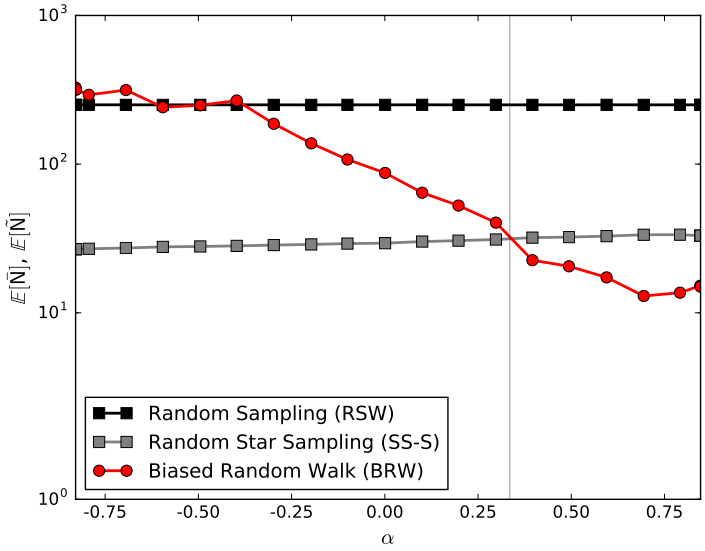

Figure 3.6: The average number of iterations to find a maximum degree node in an ER graph with $n = 500$ and $s = 0.01$ and assortativity α obtained via Alg. 1, using the biased random walk (BRW) with $\beta = 5$, denoted $\mathbb{E}[\tilde{\mathsf{N}}]$, and using random star sampling (SS-S), denoted $\mathbb{E}[\bar{\mathsf{N}}]$. $\mathbb{E}[\tilde{\mathsf{N}}]$ was calculated from 10^5 walks on 20 graphs per α_T, $\mathbb{E}[\bar{\mathsf{N}}]$ was also calculated from 10^5 trials on 20 graphs per α_T. BRW is superior to SS-S for $\alpha \geq 0.34$.

3.3.3 Local Maxima and Correlation with Biased Random Walk

The previous subsection showed that a BRW outperforms SS-S for graphs with high assortativity. This section attempts to explain this observation through a connection with the number of local maxima in the graph.

For any random undirected graph $G = (\mathcal{V}, \mathcal{E})$ define a node v as a strict local maximizer (SLM) if $d_v > \max\{d_u, u \in \Gamma_v\}$, and a non-strict local maximizer (NLM) if $d_v \geq \max\{d_u, u \in \Gamma_v\}$. Define L_S, L_M as the random number of SLM, NLM in $[n]$, respectively. Intuitively, the BRW algorithm is attempting to follow a local gradient in the graph to find a maximum degree node, and as such it selects (with high probability) one of its highest degree neighbors at each iteration. Loosely speaking, a graph with a preponderance of local maxima will not have a pronounced local gradient to follow, and as such the BRW algorithm will not perform well. The remainder of this section shows how to estimate $\mathbb{E}[\mathsf{L}_S], \mathbb{E}[\mathsf{L}_M]$, and then shows that the fraction of SLM, $\mathbb{E}[\mathsf{L}_S]/n$ correlates with the number of iterations a BRW takes to discover a maximum degree node, $\mathbb{E}[\tilde{\mathsf{N}}]$.

Recall the joint degree distribution $\mathbf{F}$ in Eq. (3.4) and define a conditional version $\mathbf{C}$ obtained by normalizing each row to sum to one, i.e., $C_{j,k} = F_{j,k}/f_j$, which may be interpreted as the probability that the "second" endpoint of an edge has degree k, conditioned on the "first" endpoint

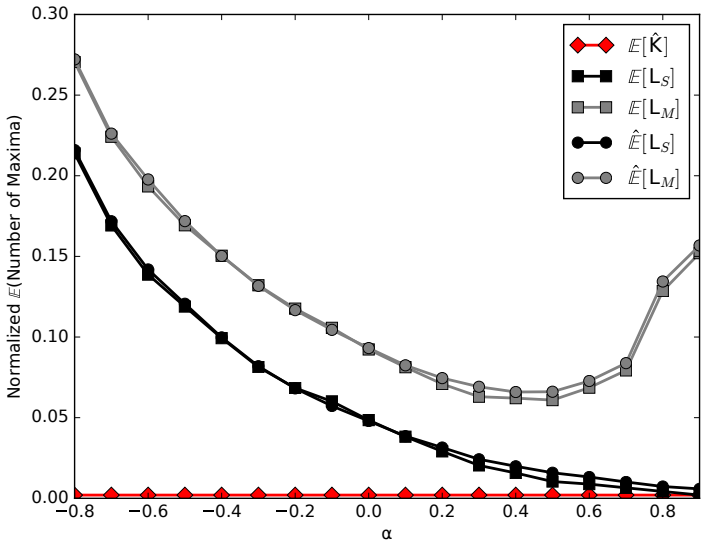

Figure 3.7: The average fraction of strict (SLM, $\mathbb{E}[\mathsf{L}_S]$) local, non-strict (NLM, $\mathbb{E}[\mathsf{L}_M]$) local, and global ($\mathbb{E}[\hat{\mathsf{K}}]$) maximizers vs. the assortativity α for ER graphs with $n = 500$ and $s = 0.01$ rewired by Alg. 1. The value predicted by Eq. (3.7) shows very close agreement with the empirical average.

having degree j. With this in hand $\mathbb{E}[\mathsf{L}_S]/n$ can be approximated as

$$\frac{\mathbb{E}[\mathsf{L}_S]}{n} \approx \sum_{j \in \bar{\mathcal{D}}} \left(\sum_{k=0}^{j-1} C_{j,k} \right)^j w_j. \tag{3.7}$$

The law of total probability used to condition on the degree j of a randomly selected non-isolated node, and the observation that such a node is a SLM iff each of its j neighbors has degree strictly less than j. The expression for NLM is identical but the sum over k has upper limit j instead of $j - 1$. The quantity is an approximation on account of treating the degrees of the j neighbors as independent. Nonetheless, the approximation is extremely accurate, as evidenced by Fig. 3.7.

Finally, Fig. 3.8 shows the correlation between the fraction of nodes that are SLM, $\mathbb{E}[\mathsf{L}_S]/n$, and the number of iterations required by the BRW, $\mathbb{E}[\tilde{\mathsf{N}}]$. The coefficient of determination, R^2, value is 0.687. If the scatter plot is restricted to the regime of $\alpha \in [1/2, 1]$, where the *absence* of SLM presumably aids the performance of the BRW, the R^2 value increases to 0.783. This demonstrates that $\mathbb{E}[\mathsf{L}_S]$ is an insightful statistic, but certainly not the only factor, in predicting $\mathbb{E}[\tilde{\mathsf{N}}]$. More work is needed to better understand the performance of a BRW as a function of α.

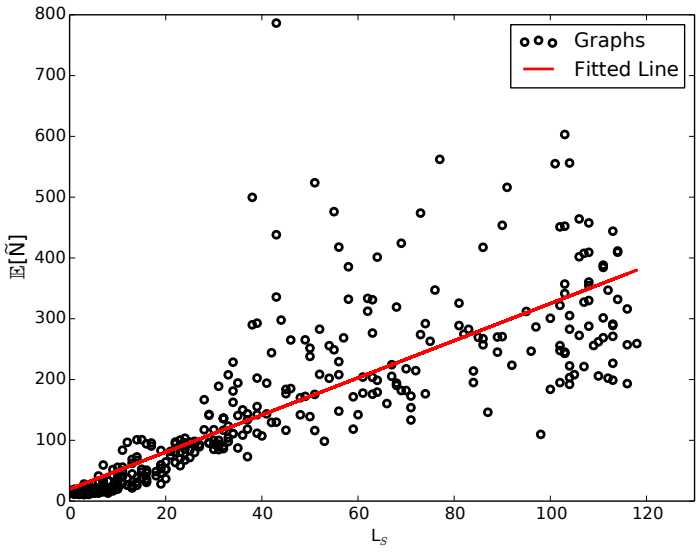

Figure 3.8: The correlation between the number of nodes that are strict local maximizers (SLM), L_S, in an ER graph with $n = 500$, $s = 0.01$ and the expected number of steps required by the biased random walk (BRW), $\mathbb{E}[\tilde{\mathsf{N}}]$, where $\beta = 5$.

3.4 Conclusions

The two questions posed in this chapter were *i*) how many maximum degree nodes are found in an ER graph, on average, and *ii*) is it better to search for these nodes via random walk or via random sampling? The answer to the first question is that there is on average a unique maximum degree node, and the answer to the second is that it depends upon the graph assortativity. Additionally this chapter establishes a connection between the performance of a biased random walk and the prevalence of strict locally maximum degree nodes in the graph, although work remains to better understand this phenomenon.

3.5 Appendix

3.5.1 Equivalence of assortativity definitions

Newman defines the assortativity α of an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with degree distribution $\mathbf{w} = (w_k, k \in \mathcal{D})$, mean degree $\mu = \sum_{k\in\mathcal{D}} k w_k$, and $\mathcal{D}$ the set of degrees, as follows. First, the excess degree is defined as $\mathbf{b} = (b_k, k \in \mathcal{B})$, for $\mathcal{B}$ be the set of excess degrees found in the graph, with $b_k = (k+1)w_{k+1}/\mu$. Observe $\mathcal{B} = \{k-1 | k \in \mathcal{D} \setminus 0\}$. Second, the joint excess degree distribution $\mathbf{B}$ is defined as having entries $B_{j,k} = B_{k,j}$, the joint distribution on the excess degrees of an edge selected at random. This quantity obeys $\sum_{j,k\in\mathcal{B}^2} B_{j,k} = 1$ and $\sum_{j\in\mathcal{B}} B_{j,k} = b_k$. Third, the excess

degree distribution $\mathbf{b} = (b_k, k \in \mathcal{B})$ has a variance $\sigma^2_{\mathbf{b}} = \sum_{k\in\mathcal{B}} k^2 b_k - (\sum_{k\in\mathcal{B}} k b_k)^2$. Finally,

$$\alpha = \frac{1}{\sigma^2_{\mathbf{b}}} \sum_{j,k\in\mathcal{D}^2} jk(B_{j,k} - b_j b_k). \tag{3.8}$$

The objective of this proof is to relate α defined in terms of $\mathbf{B}$ to the joint degree distribution, $\mathbf{F}$. The starting point is F_{jk}, the fraction of edges with degrees j, k, respectively. Let d_v be the degree of any $v \in \mathcal{V}$, and write $\hat{d}_v = d_v - 1$ for the excess degree of any non-isolated node $v \in \bar{\mathcal{V}}$, and $\hat{d}(uv) \equiv \{\hat{d}_u, \hat{d}_v\}$ for the exceess degrees of any edge $e \equiv uv \in \mathcal{E}$.

Let $A_{j,k} = A_{k,j} = \#\{e \in \mathcal{E} | \hat{d}(e) = \{j, k\}\}$ for $j, k \in \mathcal{B}$ with $j \neq k$ be the number of edges with endpoints with excess degrees j, k. Let $m_s = \#\{uv \in \mathcal{E} | \hat{d}_u = \hat{d}_v\}$ be the number of edges with endpoints of equal excess degree.

Proposition 5. *The joint excess degree distribution* $\mathbf{B}$ *is given by*

$$B_{j,k} = \begin{cases} \frac{A_{j,k}}{2m}, & j \neq k \\ \frac{A_{j,j}}{m}, & j = k \end{cases}, \tag{3.9}$$

Proof. First, observe:

$$\begin{aligned} \sum_{j,k\in\mathcal{B}} B_{j,k} &= \sum_{j,k\in\mathcal{B}^2|j\neq k} B_{j,k} + \sum_{j\in\mathcal{B}} B_{j,j} \\ &= \frac{1}{m}\left(\frac{1}{2}\sum_{j,k\in\mathcal{B}^2|j\neq k} A_{j,k} + \sum_{j\in\mathcal{B}} A_{j,j}\right) \\ &= \frac{\frac{1}{2}((m - m_s) + (m - m_s)) + m_s}{m} = 1. \end{aligned} \tag{3.10}$$

Second, $\sum_{k\in\mathcal{B}} B_{j,k} = b_j$ is proved. Observe, for any $j \in \mathcal{B}$:

$$\sum_{k\in\mathcal{B}} B_{j,k} = B_{j,j} + \sum_{k\in\mathcal{B}\setminus j} B_{j,k} = \frac{1}{m}\left(A_{j,j} + \frac{1}{2}\sum_{k\in\mathcal{B}\setminus j} A_{j,k}\right) \tag{3.11}$$

Next, recall $\sum_{v\in\mathcal{V}} d_v = 2m$, which may be written in terms of $\mathcal{D}$ as $\sum_{j\in\mathcal{D}} j n_j = 2m$, for $n_j = |\mathcal{V}_j|$ the number of nodes of degree j, with $n_j/n = w_j$. After dividing through by n, $\sum_{j\in\mathcal{D}} j w_j = \frac{2m}{n}$. With these observations it becomes clear that

$$b_j = \frac{(j+1)w_{j+1}}{\sum_{k\in\mathcal{D}} k w_k} = \frac{(j+1)n_{j+1}}{2m}. \tag{3.12}$$

Thus, the target equation $\sum_{k\in\mathcal{B}} B_{j,k} = b_j$ is equivalent, after simplifying, to:

$$2A_{j,j} + \sum_{k\in\mathcal{B}\setminus j} A_{j,k} = n_{j+1}(j+1). \tag{3.13}$$

The RHS is the number of degree $j+1$ nodes (n_{j+1}) times the number of edges ($j+1$) emanating from each such node, which is equal to the number of "edge stubs" tied to nodes of degree $j+1$. But this is the same as the LHS, which counts all edges with both endpoints are degree $j+1$ twice and edges with exactly one endpoint of degree $j+1$ once. □

Let $\bar{\mathcal{D}} = \mathcal{D} \setminus \{0\}$ be the set of non-zero degrees found in the graph. Define $K_{j,k} = K_{k,j} = \#\{e \in \mathcal{E} | d(e) = \{j,k\}\}$, for $j,k \in \mathcal{D}$. Observe $K_{j,k} = A_{j-1,k-1}$. Define the $\bar{\mathcal{D}} \times \bar{\mathcal{D}}$ symmetric matrix $\mathbf{F}$ with entries

$$F_{j,k} = \begin{cases} \frac{K_{j,k}}{2m}, & j \neq k \\ \frac{K_{j,j}}{m}, & j = k \end{cases} \tag{3.14}$$

To show the properties of $\mathbf{F}$ are analogous to that of $\mathbf{B}$, requires the overall sum and the row sum of $\mathbf{F}$.

Proposition 6. *The matrix* $\mathbf{F}$ *has the properties that*

$$\sum_{j,k\in\bar{\mathcal{D}}} F_{j,k} = 1, \ \sum_{k\in\bar{\mathcal{D}}} F_{j,k} = f_j \equiv \frac{jw_j}{\mu}, \tag{3.15}$$

where $\mathbf{f} = (f_j, j \in \bar{\mathcal{D}})$ *is a distribution.*

Proof. The proof of the first property is exactly the same as that for $\mathbf{B}$:

$$\begin{aligned} \sum_{j,k\in\bar{\mathcal{D}}} F_{j,k} &= \sum_{j,k\in\bar{\mathcal{D}}^2|j\neq k} F_{j,k} + \sum_{j\in\bar{\mathcal{D}}} F_{j,j} \\ &= \frac{1}{m}\left(\frac{1}{2}\sum_{j,k\in\bar{\mathcal{D}}^2|j\neq k} K_{j,k} + \sum_{j\in\bar{\mathcal{D}}} K_{j,j}\right) \\ &= \frac{\frac{1}{2}((m-m_s)+(m-m_s))+m_s}{m} = 1 \end{aligned} \tag{3.16}$$

The proof of the second property follows from

$$2m\sum_{k\in\bar{\mathcal{D}}} F_{j,k} = 2K_{j,j} + \sum_{k\in\bar{\mathcal{D}}\setminus j} K_{j,k} = jn_j, \tag{3.17}$$

the last equality follows from the observation that the LHS jn_j is the number of edge stubs at

degree j nodes, and this equals the RHS after recognizing edges with both endpoints having degree j consume two stubs each. Finally, observe:

$$\sum_{k\in\mathcal{D}_0} F_{j,k} = \frac{jn_j}{2m} = \frac{jn_j}{\sum_k kn_k} = \frac{jw_j}{\sum_k kw_k} = \frac{jw_j}{\mu} = a_j. \tag{3.18}$$

□

Next assortativity α is expressed in terms of $\mathbf{F}$ and $\mathbf{f}$ instead of $\mathbf{B}$ and $\mathbf{b}$.

Proposition 7. *The assortativity α may be expressed in terms of $\mathbf{F}$ and $\mathbf{f}$ as*

$$\alpha = \frac{1}{\sigma^2_{\mathbf{f}}} \sum_{j,k\in\bar{\mathcal{D}}} jk(F_{j,k} - f_j f_k) \in [-1, +1]. \tag{3.19}$$

Proof. Use the change of variables $s = j + 1$ and $t = k + 1$, and observe summing over all $j, k \in \mathcal{B}$ is the same as summing over all $s, t \in \bar{\mathcal{D}}$. Second, observe $C_{jk} = \hat{C}_{j-1,k-1}$ means $F_{jk} = B_{j-1,k-1}$, and $b_k = (k+1)w_{k+1}/\mu$ and $f_k = kw_k/\mu$ means $f_k = b_{k-1}$. Third, define the variance of $\mathbf{f}$ as

$$\sigma^2_{\mathbf{f}} = \sum_{k\in\bar{\mathcal{D}}} k^2 f_k - \left(\sum_{k\in\bar{\mathcal{D}}} kf_k\right)^2, \tag{3.20}$$

and observe

$$\sigma^2_{\mathbf{b}} = \sum_{k\in\mathcal{B}} k^2 b_k - \left(\sum_{k\in\mathcal{B}} kb_k\right)^2 = \sum_{c\in\bar{\mathcal{D}}} (c-1)^2 f_c - \left(\sum_{c\in\bar{\mathcal{D}}} (c-1) f_c\right)^2 = \sigma^2_{\mathbf{f}}. \tag{3.21}$$

These three observations allow

$$\begin{aligned} \alpha &= \frac{1}{\sigma^2_{\mathbf{b}}} \sum_{j,k\in\mathcal{B}} jk(B_{j,k} - b_j b_k) \\ &= \frac{1}{\sigma^2_{\mathbf{f}}} \sum_{j,k\in\bar{\mathcal{D}}} (j-1)(k-1)(B_{j-1,k-1} - b_{j-1}b_{k-1}) \\ &= \frac{1}{\sigma^2_{\mathbf{f}}} \sum_{j,k\in\bar{\mathcal{D}}} (j-1)(k-1)(F_{j,k} - f_j f_k) \end{aligned} \tag{3.22}$$

Simple algebra shows the $(j-1)(k-1)$ in the summand may be replaced with jk without affecting the sum. □

3.5.2 Proof of theorem 2

The following elementary lemma is found as equation (2.5) in Brands et al. [1994], and is used in the proof of Prop. 4.

Lemma 3. *For $0 < b < a$ and $n \in \mathbb{N}$: $n(a-b)b^{n-1} \leq a^n - b^n \leq n(a-b)a^{n-1}$.*

Proof. Observe $a^n - b^n = (a-b)\sum_{j=0}^{n-1} a^j b^{n-1-j}$. The bounds follows from $b^{n-1} < a^j b^{n-1-j} < a^{n-1}$. □

Proof of Prop. 4. i) Proof of $\mathbb{E}[\mathsf{K}(n)] \leq u(k,n)$. Split the sum in Lem. 1 at $k \in [n]$, apply the definition of r_{j+1} in the first term and R_j in the second term, use the facts that $\mathbf{r}$ is increasing and $\mathbf{R}$ is decreasing to pull out r_k and R_k^{n-1}, apply Lem. 3 to each term in each sum, and telescope the two sums to yield $u(k,n)$:

$$\begin{aligned}
\mathbb{E}[\mathsf{K}] &= \sum_{j=0}^{k-1} n w_j W_j^{n-1} + \sum_{j=k}^{n-1} n w_j W_j^{n-1} \\
&= \sum_{j=0}^{k-1} n r_{j+1} w_{j+1} W_j^{n-1} + \sum_{j=k}^{n-1} n w_j R_j^{n-1} W_{j-1}^{n-1} \\
&\leq r_k \sum_{j=0}^{k-1} n w_{j+1} W_j^{n-1} + R_k^{n-1} \sum_{j=k}^{n-1} n w_j W_{j-1}^{n-1} \\
&\leq r_k \sum_{j=0}^{k-1} (W_{j+1}^n - W_j^n) + R_k^{n-1} \sum_{j=k}^{n-1} (W_j^n - W_{j-1}^n) \qquad (3.23)
\end{aligned}$$

ii) Proof of $u(k,n) \leq v(k,n)$. Observe $W_k^n - w_0^n \leq 1$, $R_k^{n-1} \leq R_k^n$, and $1 - W_{k-1}^n \leq n(1 - W_{k-1})$ via Lem. 3. □

The proof of Lem. 2, below, relies upon the following well-known bounds on the binomial CDF, found in Blake and Darabian [1987].

Lemma 4. *For the binomial distribution: i) $w_k/W_k > 1 - r_k$ for all k; ii) $1 - W_{k-1} < \frac{r_{k+1}}{r_{k+1}-1} w_k$ for k with $r_k > 1$.*

Proof of Lem. 2. i) Proof that $\mathbf{r}$ is increasing. By definition, for $k \in [n-1]$ and $\varrho \equiv (1-s)/s$:

$$r_k = \frac{w_{k-1}}{w_k} = \frac{\binom{n-1}{k-1} s^{k-1}(1-s)^{n-k}}{\binom{n-1}{k} s^k (1-s)^{n-1-k}} = \frac{k}{n-k}\varrho. \qquad (3.24)$$

This sequence is increasing, as evident from

$$\frac{1}{\varrho}(r_{k+1} - r_k) = \frac{n}{(n-k)(n-(k+1))} > 0. \tag{3.25}$$

Solving $r_k = 1$ gives $k = ns$, which, if ns is not an integer, means $r_{\lfloor ns \rfloor} \leq 1 \leq r_{\lceil ns \rceil}$.

ii) Proof that $\mathbf{R}$ is decreasing. Consider separately *a*) $k \geq \lceil ns \rceil$ ($r_k > 1$) and *b*) $k \leq \lfloor ns \rfloor$ ($r_k \geq 1$). First, for $k \geq \lceil ns \rceil$, $w_{k-1} > p_k$ and as such the expression $(w_{k-1} - w_k)W_{k-1} > -w_{k-1}w_k$ is valid for all such k, and this expression may be rearranged as

$$-w_{k-1}W_{k-1} + w_k W_{k-1} - w_{k-1}w_k < 0. \tag{3.26}$$

Second, for $k \leq \lfloor ns \rfloor$, $w_k > w_{k-1}$ and as such the expression $(w_k - w_{k-1})W_{k-1} < w_{k-1}w_k$ (*) may also be arranged as Eq. (3.26). But (*) may also be written as $W_{k-1} < \frac{w_{k-1}w_k}{w_k - w_{k-1}}$, and by adding w_k to both sides, as $W_k < w_k/(1-w_k)$, which is true for all such k by Lem. 4. Thus Eq. (3.26) holds for all $k \in [n-1]$. Now observe the equivalence

$$R_{k-1} > R_k \Leftrightarrow \frac{W_{k-1}}{W_{k-1} - w_{k-1}} > \frac{W_{k-1} + w_k}{W_{k-1}}. \tag{3.27}$$

Cross multiplying and simplifying gives Eq. (3.26). □

The proof of Thm. 2 below makes use of the following elementary upper bound on the binomial PMF w_k, which employs an upper bound on the binomial coefficient.

Lemma 5. *The binomial PMF w_k has upper bound $\tilde{w}_k \equiv$*

$$\frac{1}{\sqrt{2\pi k\left(1 - \frac{k}{n-1}\right)}}\left(\frac{(n-1)s}{k}\right)^k \left(\frac{(n-1)(1-s)}{n-1-k}\right)^{n-1-k}. \tag{3.28}$$

At $k = t(n-1)$ the upper bound is

$$w_{t(n-1)} \leq \tilde{w}_{t(n-1)} = \frac{h(t)^{n-1}}{\sqrt{2\pi t(1-t)(n-1)}}, \tag{3.29}$$

for

$$h(t) \equiv \left(\frac{s}{t}\right)^t \left(\frac{1-s}{1-t}\right)^{1-t}. \tag{3.30}$$

Proof of Lem. 5. Reparameterize the binomial coefficient upper bound of Corollary 1 in Sasvári [1999] into an upper bound on $\binom{n-1}{k}$, substitute into the binomial PMF w_k and simplify. The

expression for $\tilde{w}_{t(n-1)}$ is simple algebra. □

Proof of Thm. 2. i) Proof that $\mathbb{E}[\mathsf{K}(n)] \leq y(t(n-1), n)$ for sufficiently large n and $t \in (s, 1)$. Writing $1-W_{k-1} \leq b(k,n)$ and $R_k \leq 1+a(k,n)$, substitution into Prop. 4, and application of the exponential bound gives

$$\begin{aligned} v(k,n) &\leq r_k + nb(k,n)\left(1+\frac{na(k,n)}{n}\right)^n \\ &\leq r_k + nb(k,n)\mathrm{e}^{na(k,n)} \end{aligned} \tag{3.31}$$

Then $\mathbb{E}[\mathsf{K}(n)] \leq y(t(n-1), n)$ holds for sufficiently large n and $t \in (s, 1)$, provided $a(t(n-1), n) \leq b(t(n-1), n)$ for the same, which is show below.

First, $b(k,n) = \frac{r_{k+1}}{r_{k+1}-1} w_k$, as stated in the theorem, directly by Lem. 4. Next, also by Lem. 4,

$$R_k = 1 + \frac{w_k}{W_{k-1}} \leq 1 + \frac{w_k}{1-b(k,n)}, \tag{3.32}$$

and as such $a(k,n) = w_k/(1-b(k,n))$. Rearranging gives

$$a(k,n) \leq b(k,n) \Leftrightarrow w_k \leq q_k \equiv \frac{r_{k+1}-1}{r_{k+1}^2}. \tag{3.33}$$

Consider first q_k. By Eq. (3.24)

$$q_k = \frac{(k+1-ns)(n-(k+1))s}{((k+1)(1+s))^2}, \tag{3.34}$$

which at $k = t(n-1)$ gives

$$q_{t(n-1)} = \frac{(t(n-1)+1-ns)(n-(t(n-1)+1))s}{((t(n-1)+1)(1+s))^2}. \tag{3.35}$$

Dividing numerator and denominator by n^2 makes clear that

$$\lim_{n\to\infty} q_{t(n-1)} = \frac{(t-s)(1-t)s}{(t(1+s))^2} > 0, \tag{3.36}$$

with the inequality valid since $t \in (s, 1)$. Next consider $w_{t(n-1)} \leq \tilde{w}_{t(n-1)}$ from Lem. 5, and in particular observe

$$\frac{\partial}{\partial t} h(t) = h(t) \log \frac{s(1-t)}{(1-s)t} < 0, \tag{3.37}$$

the latter inequality on account of $\frac{s(1-t)}{(1-s)t} \in (0, 1)$ for $t \in (s, 1)$. It follows that $h(t) \leq h(s) = 1$ over

$t \in (s,1)$, and thus $h(s,t) < 1$ over $t \in (s,1)$. Consequently, $\tilde{w}_{t(n-1)} \to 0$, and therefore $w_{t(n-1)} \to 0$. It follows from Eq. (3.33) that $a(t(n-1),n) \leq b(t(n-1),n)$ for $t \in (s,1)$ for sufficiently large n.

ii) Proof that $y(t(n-1),n) \to y(t)$ as $n \to \infty$. By the continuity of the function xe^x in $y(m,n)$, it suffices to show $nb(t(n-1),n) \to 0$ and that $r_{t(n-1)} \to y(t)$ with $y(t)$ given in the theorem. First, observe $nb(t(n-1),n)$ may be upper bounded as

$$nb(t(n-1),n) \leq \frac{r_{t(n-1)+1}}{r_{t(n-1)+1}-1} n\tilde{w}_{t(n-1)}, \tag{3.38}$$

using $b(k,n) = \frac{r_{k+1}}{r_{k+1}-1} w_k$ and $w_k \leq \tilde{w}_k$. This bound converges to 0 as is now show. First, compute

$$\begin{aligned} \frac{r_{k+1}}{r_{k+1}-1} &= \frac{(k+1)(1-s)}{k+1-ns} \\ \frac{r_{t(n-1)+1}}{r_{t(n-1)+1}-1} &= \frac{(t(n-1)+1)(1-s)}{t(n-1)+1-ns}. \end{aligned} \tag{3.39}$$

Dividing the numerator and denominator of the latter expression by n and taking the limit yields $r_{t(n-1)+1}/(r_{t(n-1)+1}-1) \to t(1-s)/(t-s)$. Second, the same argument used above to show $\tilde{w}_{t(n-1)} \to 0$ also shows $n\tilde{w}_{t(n-1)} \to 0$, and as such $nb(t(n-1),n) \to 0$. It follows that $\lim_{n\to\infty} y(t(n-1),n) = \lim_{n\to\infty} r_{t(n-1)}$, and the latter equals $y(t)$.

iii) Proof that $\lim_{n\to\infty} \mathbb{E}[\mathsf{K}(n)] \leq 1+\epsilon$ for arbitrary $\epsilon > 0$. Having established $\mathbb{E}[\mathsf{K}(n)] \leq y(t(n-1),n)$ for $t \in (s,1)$ and that $y(t(n-1),n) \to y(t) = \frac{t(1-s)}{(1-t)s}$ for any $t \in (s,1)$. Observe

$$\frac{\partial}{\partial t} y(t) = \frac{y(t)}{t(1-t)} > 0 \tag{3.40}$$

and as such $y(t) \geq y(s)$ over $t \in (s,1)$. The equation $1+\epsilon = y(t)$ yields $t = \frac{s(1+\epsilon)}{1+s\epsilon} \in (s,1)$ at which value $\lim_{n\to\infty} \mathbb{E}[\mathsf{K}(n)] \leq 1+\epsilon$, as required.

Finally, to reiterate Rem. 1, observe $y(t)$ is increasing in t, with the tightest bound on the asymptotic value of $\mathbb{E}[\mathsf{K}(n)]$ achieved as $t \downarrow s$, while $h(t)$ is decreasing in t, with the consequence that the rate of convergence of $\tilde{w}_{t(n-1)}$ to zero is decreasing as $t \downarrow s$. □

Chapter 4: The self-avoiding walk-jump (SAWJ) algorithm for finding maximum degree nodes in large graphs

4.1 Introduction

4.1.1 Motivation

Many large datasets are in the form of *graphs*, e.g., *i*) social networks like Facebook or LinkedIn, *ii*) community structures like arXiv paper authorship or IMDB, or *iii*) system diagrams like protein networks or the electrical grid. In graph's edges denote influence or connection among nodes, hence maximum degree nodes are, in a sense, the most influential or most highly connected. It is a natural objective to seek out maximum degree nodes, and for large graphs this must be done by a search or sampling algorithm. The distinction between finding (or estimating) the maximum degree and finding a maximum (or near-maximum) degree node is emphasized here as this chapter focuses on finding a maximum degree node.

In this chapter it is *assumed* that the maximum degree, denoted λ, is known *a priori*, and so the problem is to find any member of the set of maximum degree nodes, denoted $\mathcal{V}_\lambda$. Although this assumption appears artificial, it is not substantially different from the more practical scenario where λ is unknown and the time available for the search is bounded. That is, the same algorithm may well be used to *i*) find the largest degree possible, without knowledge of λ, over a bounded search time, and *ii*) knowing λ, seek to find any node in $\mathcal{V}_\lambda$, in as short a time as possible.

It is also natural to consider the problem of finding *all* the maximum degree nodes, i.e., the entire set $\mathcal{V}_\lambda$, instead of just any one member of $\mathcal{V}_\lambda$. Interestingly, as shown in Chap. 3, for an important class of random graphs, with high probability there is a unique maximum degree node, i.e., $|\mathcal{V}_\lambda| \approx 1$. Thus algorithms for the problem of finding one maximizer will also apply to the problem of finding all maximizers.

Without knowing the graph structure, the natural approach to finding a max degree node is to sample. A prominent approach to sampling graphs is "star sampling", wherein a sample consists of a node u and its immediate neighborhood, denoted Γ_u. While this approach is reasonable for some graphs, it is suboptimal in (many) others, precisely because it ignores *local information*, i.e., the degrees of the neighbors of the current node. Exploiting this local information is the focus of random walk graph search algorithms, which repeatedly select, often at random and with bias based on neighbor degrees, a neighbor of the current node.

The main idea in this chapter is to combine these two fundamental approaches (random sampling and random walks) into a single algorithm called a *self-avoiding walk-jump (SAWJ)* algorithm sometimes abbreviated as SJ. SAWJ exploits local information by repeatedly following maximum degree neighbors, as in random *walk*, or *jumps* to a randomly selected node, as in sampling, when the local information fails to provide a path towards high degree nodes. SAWJ is particularly effective on so-called *assortative* graphs, where $\alpha > 0$ for $\alpha \in [-1, +1]$ is defined as the Pearson correlation coefficient of the degrees of the endpoints of a randomly selected edge.

4.1.2 Related work

The literature on random walks on graphs for estimating graph properties is too large to cover; however here are a few highlights that were influential in the development of SAWJ. Avin and Krishnamachari [2008] was one of the first to propose random walks biased towards unvisited high degree nodes. Lee et al. [2012] proposed a self avoiding random walk with edge re-weighting. The Albatross Sampling (AL) algorithm introduced by Jin et al. [2011] is a Metropolis Hastings random walk with random vertex sampling. The Frontier Sampling (FS) algorithm introduced by Ribeiro and Towsley [2010] employs multiple walkers that visit nodes with the same frequency as a simple random walk. Yet collectively the walkers achieve a lower mean-squared error in estimating graph properties than a simple random walk.

Recently Avrachenkov introduced two algorithms that operate in the same problem space as SAWJ. First Avrachenkov introduced an algorithm Avra-W (AW) similar to AL which uses uniform restarts to find the top k degree nodes with high probability Avrachenkov et al. [2012]. Second Avrachenkov introduced Avra-S (AS) which introduces a random sampling scheme to find high degree nodes Avrachenkov et al. [2014]. A key difference between SAWJ and AW is that SAWJ does not have a uniform restart probability. SAWJ's restart probability depends on its degree neighborhood and the assortativity of the graph being searched. In the case of AS, Avrachenchov assumes the cost of sampling the degree neighborhood of a node scales with the neighborhoods size. Counter to AS, SAWJ assumes that the cost of sampling a nodes degree neighborhood is fixed in the neighborhoods size. The differing assumptions of AS and SAWJ about the underlying cost model lead to the development of very different algorithms despite the fact that Avrachenchov's algorithms and SAWJ solve the same problem. This chapter compares AL, FS, AW, and AS against SAWJ under the assumption of fixed or unit cost in sampling a nodes degree neighborhood and linear cost, which scales with to the size of a nodes neighborhood. The first assumption is called the the *unit cost model* and the second the *linear cost model*.

4.1.3 Contributions

The key contribution of this chapter is the introduction of the SAWJ algorithm. This chapter is broken down into three braod sections, first is given a mathematical model to approximate the expected search time of a simplified version of SAWJ, entitled Walk Jump (WJ), using the theory of absorbing Markov chains. Second is introduced the SAWJ algorithm. Third, the performance of SAWJ is evaluated relative to WJ, the two variants of star sampling, SS-R and SS-S, Albatross Sampling (AS) Jin et al. [2011], Frontier Sampling (FS) Ribeiro and Towsley [2010], Avria-W, and Avria-S on a class of graphs with tunable assortativity, called assortative Erdős-Rényi (AER) random graphs. This chapter is organized as follows. The notation and definition are given in Sec. 4.2 AER graphs and the maximum neighborhood degree distribution Sec. 4.3 and Sec. 4.4 gives a Markov chain model for the steps to find a maximum degree node with WJ. The SAWJ algorithm and its performance relative to the competing algorithms under the unit and linear cost model are presented in Sec. 4.5. A brief conclusion is given in Sec. 4.6 and Sec. 4.7 is the appendix.

4.2 Notation and Definitions

4.2.1 Notation conventions

We write $[n]$ to denote $\{0, \ldots, n-1\}$, for $n \in \mathbb{N}$. Sets are denoted in a calligraphic font, e.g., $\mathcal{D}, \mathcal{E}, \mathcal{V}$, etc., and often sets are partitioned into subsets with either single or double index, e.g., $\mathcal{E}_j, \mathcal{E}_{j,k}$. Collections of subsets are denoted in a script font, e.g., $\mathscr{E} = (\mathcal{E}_j, j \in \mathcal{D})$ denotes the collection of subsets $\mathscr{E}$ with elements (subsets) $\mathcal{E}_j$ indexed by j with index set $\mathcal{D}$. Quantities with a single (double) index are treated as a vector (matrix) and are denoted by a bold lower- (upper-) case roman symbol, e.g., $\mathbf{m} = (m_j, j \in \mathcal{D})$, and $\mathbf{M} = (M_{j,k}, (j,k) \in \mathcal{D}^2)$.

Random variables (RVs) are denoted in sans-serif font, $\mathsf{x}, \mathsf{y}, \mathsf{X}, \mathsf{Y}$, etc., with expectation ($\mathbb{E}[\cdot]$), probability ($\mathbb{P}(\cdot)$), covariance ($\mathrm{Cov}(\cdot,\cdot)$), correlation ($\mathrm{Corr}(\cdot,\cdot)$), variance ($\mathrm{Var}(\cdot)$), and standard deviation ($\mathrm{Std}(\cdot)$). Write $\mathsf{u} \sim \mathrm{Uni}(\mathcal{V})$ to denote a discrete uniform RV drawn from $\mathcal{V}$, $\mathsf{X} \sim \mathrm{Bin}(n, s)$ to denote a binomial RV with parameters (n, s), $\mathsf{Z} \sim \mathcal{N}(\mu_0, \sigma_0^2)$ to denote a normal RV with mean μ_0 and variance σ_0^2, and let $Q(\cdot)$ denote the standard normal CDF, i.e., $Q(z) \equiv \mathbb{P}(\mathsf{Z} \leq z)$ for $\mathsf{Z} \sim \mathcal{N}(0,1)$. This chapter uses the acronyms PDF/PMF for probability density/mass function, CDF for cumulative distribution function, and IID for independent and identically distributed.

Consider an undirected simple graph $G = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = n$ nodes and $|\mathcal{E}| = m$ edges. The set of edges is a collection of unordered pairs from $\mathcal{V}$, with $uv \in \mathcal{E}$ denoting an edge connecting nodes u and v. The *degree* of node u is the number of u's neighbors: $d_u \equiv |\Gamma(u)|$, where $\Gamma_u \equiv \{v \in \mathcal{V} | uv \in \mathcal{E}\}$

is the neighborhood of u. Denote the set of maximum degree neighbors of u as $\Gamma_u^+ := \text{argmax}_{v\in\Gamma_u} d_v$, and the maximum degree of a neighbor of u as $\lambda_u = d_v$ for $v \in \Gamma_u^+$. Let $\mathcal{D} \equiv \{d_v | v \in \mathcal{V}\}$ denote the *set of degrees* found in the graph and $\mathbf{d}$ be the degree sequence of G. Partition the nodes $\mathcal{V}$ by degree into a collection of subsets $\mathscr{V} \equiv (\mathcal{V}_k, k \in \mathcal{D})$, with $\mathcal{V}_k \equiv \{v \in \mathcal{V} | d_v = k\}$ the nodes of degree k. The *degree distribution* is denoted by $\mathbf{w} \equiv (w_k, k \in \mathcal{D})$, with $w_k \equiv |\mathcal{V}_k|/n$ the fraction of nodes of degree k. The *average degree*, over $v \in \mathcal{V}$, is $\mu \equiv \sum_{k\in\mathcal{D}} k w_k = 2m/n$ and the degree variance is denoted by σ^2. Let $\lambda \equiv \max_{v\in\mathcal{V}} d_v = \max(\mathcal{D})$ be the *maximum degree*, $\mathcal{V}_\lambda \equiv \text{argmax}_{v\in\mathcal{V}} d_v$ be the set of *maximum degree nodes*, ϕ be the diameter of G, and $\blacktriangle$ the fraction of triangles in G.

The set of isolated nodes are those with zero degree $\mathcal{V}_0$. Non-isolated (henceforth NI) nodes are captured as follows. Let $\bar{\mathcal{V}} \equiv \mathcal{V} \setminus \mathcal{V}_0 = \{v \in \mathcal{V} | d(v) > 0\}$ be the *NI nodes*, $\bar{n} \equiv |\bar{\mathcal{V}}|$ the number of such nodes, and $\bar{\mathcal{D}} \equiv \mathcal{D} \setminus \{0\}$ the set of degrees found among NI nodes. Define the *NI degree distribution* $\bar{\mathbf{w}} \equiv (\bar{w}_k, k \in \bar{\mathcal{D}})$ with entries $\bar{w}_k \equiv |\mathcal{V}_k|/\bar{n}$. The *average NI degree*, over $v \in \bar{\mathcal{V}}$, is $\bar{\mu} \equiv \sum_{k\in\bar{\mathcal{D}}} k\bar{w}_k = 2m/\bar{n}$.

4.2.2 Joint and marginal degree distribution

Partition the edges $\mathcal{E}$ into the collection of subsets $\mathscr{E} = (\mathcal{E}_{j,k}, (j,k) \in \bar{\mathcal{D}}^2)$ with elements

$$\begin{aligned} \mathcal{E}_{j,k} &\equiv \{uv \in \mathcal{E} | \{d_u, d_v\} = \{j,k\}\},\ j \neq k \\ \mathcal{E}_{j,j} &\equiv \{uv \in \mathcal{E} | d_u = d_v = j\} \end{aligned} \tag{4.1}$$

Here $\mathcal{E}_{jk}$ are edges with degrees $\{j,k\}$, and $\mathcal{E}_{j,j}$ are edges with both endpoints of degree j. Define the *symmetric* $|\bar{\mathcal{D}}| \times |\bar{\mathcal{D}}|$ *joint degree distribution* matrix $\mathbf{F}$ with entries F_{jk}:

$$F_{jk} \equiv \begin{cases} \frac{|\mathcal{E}_{j,k}|}{2m}, & j \neq k \\ \frac{|\mathcal{E}_{j,j}|}{m}, & j = k \end{cases} \tag{4.2}$$

and observe $\sum_{(j,k)\in\bar{\mathcal{D}}^2} F_{j,k} = 1$. Also let the collection $\mathscr{E} \equiv (\mathcal{E}_k, k \in \bar{\mathcal{D}})$ with elements $\mathcal{E}_k \equiv \{uv \in \mathcal{E} | d_u = k \text{ or } d_v = k\}$ denote the set of edges with *one or both* endpoints of degree k.

The marginal degree distribution $\mathbf{f} \equiv (f_k, k \in \bar{\mathcal{D}})$ corresponding to the joint degree distribution $\mathbf{F}$ is comprised of the row sums from $\mathbf{F}$:

$$f_k \equiv \sum_{j\in\bar{\mathcal{D}}} F_{j,k} = \frac{|\mathcal{E}_k| - |\mathcal{E}_{k,k}|}{2m} + \frac{|\mathcal{E}_{k,k}|}{m},\ k \in \bar{\mathcal{D}}. \tag{4.3}$$

Here f_k is the probability that a randomly chosen endpoint of an edge selected uniformly at random

from $\mathcal{E}$ has degree k. It is a standard result that

$$f_k = \frac{k\bar{w}_k}{\sum_{j\in\bar{\mathcal{D}}} j\bar{w}_j},\ k \in \bar{\mathcal{D}}, \tag{4.4}$$

meaning that selecting a random vertex by first selecting a random edge biases the selection towards vertices of higher degree. To see Eq. (4.4), observe the following two quantities are equal: *i*) Eq. (4.4) equals $\frac{k|\mathcal{V}_k|}{2m}$, and *ii*) Eq. (4.3) equals $\frac{|\mathcal{E}_k|+|\mathcal{E}_{k,k}|}{2m}$, and both expressions equal the fraction of the $2m$ edge "stubs" connected to a node of degree k. The corresponding mean and variance of $\mathbf{f}$ are given by the usual formulae:

$$\mu_{\mathbf{f}} \equiv \sum_{k\in\bar{\mathcal{D}}} k f_k,\ \sigma^2_{\mathbf{f}} \equiv \sum_{k\in\bar{\mathcal{D}}} (k-\mu_{\mathbf{f}})^2 f_k. \tag{4.5}$$

4.2.3 Max neighbor degree distribution and degree assortativity

For any NI node $u \in \bar{\mathcal{V}}$, define $\lambda_u \equiv \max_{v\in\Gamma_u}(d_v)$ as the maximum degree over the neighbors of u, hereafter called the *maximum neighbor degree*. Define the set of nodes with degree j and maximum neighbor degree k as

$$\mathcal{V}_{j,k} \equiv \{u \in \mathcal{V} | d_u = j, \lambda_u = k\},\ (j,k) \in \bar{\mathcal{D}}^2. \tag{4.6}$$

Define the $|\bar{\mathcal{D}}| \times |\bar{\mathcal{D}}|$ conditional distribution matrix $\mathbf{H}$ with entries

$$H_{j,k} \equiv \frac{|\mathcal{V}_{j,k}|}{|\mathcal{V}_j|},\ (j,k) \in \bar{\mathcal{D}}^2, \tag{4.7}$$

where $H_{j,k}$ is the probability that a randomly selected NI node of degree j has maximum neighbor degree k. Observe $\sum_{k\in\bar{\mathcal{D}}} H_{j,k} = 1$ for each $j \in \bar{\mathcal{D}}$. Let $\mathsf{Y} = d_{\mathsf{u}}$ be the random degree of a randomly selected NI node $\mathsf{u} \sim \text{Uni}(\bar{\mathcal{V}})$, and let $\mathsf{M} = \lambda_{\mathsf{u}}$ be the corresponding maximum neighbor degree. Then $\mathbb{P}(\mathsf{M} = k|\mathsf{Y} = j) = H_{j,k}$.

Fix a graph $G = (\mathcal{V}, \mathcal{E})$, and let (X, Y) be the degrees of a randomly selected edge.

Definition 9. *Graph* assortativity[1] *is the Pearson correlation coefficient of the degrees of a randomly selected edge,* $\alpha \equiv \rho_{\mathsf{X},\mathsf{Y}}$.

Since X, Y are identically distributed:

$$\alpha \equiv \frac{\mathbb{E}[\mathsf{XY}] - \mathbb{E}[\mathsf{X}]^2}{\text{Var}(\mathsf{X})} = \frac{1}{\sigma^2_{\mathbf{f}}}\left(\sum_{(j,k)\in\bar{\mathcal{D}}^2} jkF_{j,k} - \mu^2_{\mathbf{f}}\right). \tag{4.8}$$

[1]Note, the degree assortativity of a graph was first defined by Newman [2002] in terms of the *joint excess degree distribution*; it is straightforward to see that the definition in terms of the *joint degree distribution* $\mathbf{F}$ is equivalent, Chap. 3.

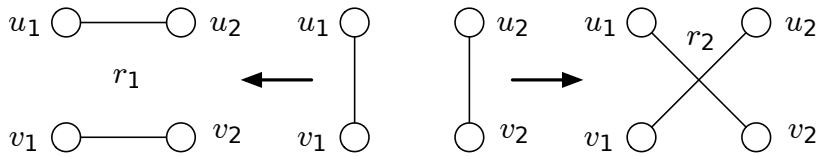

Figure 4.1: Illustration of the primitive rewiring operation in Def. 10. The two edges u_1v_1 and u_2v_2 (middle) are replaced either with u_1u_2 and v_1v_2 (rewiring r_1, left) or with u_1v_2 and u_2v_1 (rewiring r_2, right), provided neither of the new edges are already present. Rewiring leaves all degrees unchanged.

The following lemma offers insight into the above sum.

Lemma 6. *Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with $|\mathcal{E}| = m$, NI degrees $\bar{\mathcal{D}}$, and joint degree distribution $\mathbf{F}$. Then*

$$\sum_{(j,k)\in\bar{\mathcal{D}}^2} jkF_{j,k} = \mathbb{E}[\mathsf{XY}] = \frac{1}{m}\sum_{uv\in\mathcal{E}} d_u d_v. \tag{4.9}$$

Here (X, Y) are the degrees of a randomly selected edge, and the right side is the product of the degrees of the endpoints of an edge, averaged over all edges. The lemma follows by substituting the appropriate definitions.

4.2.4 Degree preserving rewiring

Define the following primitive rewiring operation on a graph $G = (\mathcal{V}, \mathcal{E})$, also used in Li et al. [2005] and in Chap. 3.

Definition 10. *The primitive rewiring operation replaces any two edges u_1v_1 and u_2v_2 from $\mathcal{E}$ with either i) u_1u_2 and v_1v_2 (provided neither of these edges are already in $\mathcal{E}$) or ii) u_1v_2 and u_2v_1 (provided neither of these edges are already in $\mathcal{E}$):*

$$\begin{aligned} \mathcal{E} &\rightarrow r_1(\mathcal{E}) \equiv \{\mathcal{E} \setminus \{u_1v_1, u_2v_2\}\} \cup \{u_1u_2, v_1v_2\} \\ \mathcal{E} &\rightarrow r_2(\mathcal{E}) \equiv \{\mathcal{E} \setminus \{u_1v_1, u_2v_2\}\} \cup \{u_1v_2, u_2v_1\} \end{aligned} \tag{4.10}$$

These operations are illustrated in Fig. 4.1. The operation is degree preserving, i.e., the degree of each node is unchanged by rewiring. Let $G = (\mathcal{V}, \mathcal{E})$ be the graph before the operation, fix edges u_1v_1 and u_2v_2, and let $G_1 = G(\mathcal{V}, r_1(\mathcal{E}))$ and $G_2 = G(\mathcal{V}, r_2(\mathcal{E}))$ denote the two possible graphs after the operation. Let $\alpha(G), \alpha(G_1), \alpha(G_2)$ denote the corresponding assortativities. The following lemma computes the impact of the primitive rewiring operation on the assortativity.

Lemma 7. *Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with $|\mathcal{E}| = m$, and fix two edges u_1v_1 and u_2v_2. Then the changes in assortativity under the two possible rewirings are $\Delta_i \equiv \alpha(G_i) - \alpha(G)$, for $i \in \{1, 2\}$, where*

$$\begin{aligned} \Delta_1 &= \frac{1}{m}\left(d_{u_1}d_{u_2} + d_{v_1}d_{v_2} - C\right) \\ \Delta_2 &= \frac{1}{m}\left(d_{u_1}d_{v_2} + d_{u_2}d_{v_1} - C\right) \end{aligned} \tag{4.11}$$

and $C \equiv d_{u_1}d_{v_1} + d_{u_2}d_{v_2}$.

Proof. The impact of rewiring on $\alpha(G_i)$ in Eq. (4.8) is restricted to the summation, by virtue of the fact that the rewiring is degree preserving, and therefore μ, σ^2 are unaffected. By Lem. 6, the only edges that change under the rewiring are the two removed edges and the two added edges. □

The term $d_{u_1}d_{v_1} + d_{u_2}d_{v_2}$ reflects the two removed edges, while $d_{u_1}d_{u_2} + d_{v_1}d_{v_2}$ and $d_{u_1}d_{v_2} + d_{u_2}d_{v_1}$ reflect the two possible pairs of new edges.

4.3 Assortative Erdős Rényi (AER) Graphs

A (non-assortative) Erdős-Rényi (ER) random graph $G_\epsilon = (\mathcal{V}, \mathcal{E})$ has parameters (n, s), with $n \in \mathbb{N}$ and $s \in (0, 1)$. The node set is $\mathcal{V} = [n]$, and the *random* edge set $\mathcal{E}$ is formed by adding each of the $\binom{n}{2}$ possible edges independently with probability s. Let $(\mathsf{X}_u, u \in \mathcal{V})$ be the random degrees in G_ϵ; by construction it is clear that $\mathsf{X}_u \sim \text{Bin}(n-1, s)$ for each $u \in \mathcal{V}$. Conditioned on $uv \in \mathcal{E}$, their random degrees (X, Y) each have distribution $1 + \text{Bin}(n-2, s)$, where each binomial RV is formed from a separate set of (independent) $n-2$ Bernoulli RVs, and as such (X, Y) are (conditionally) independent. Thus the expected assortativity α for an ER graph $G_\epsilon = (\mathcal{V}, \mathcal{E})$ is 0.

4.3.1 Definition of an AER graph

Definition 11. *An assortative Erdős-Rényi (AER) random graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ has parameters (n, s, α), with (n, s) as defined for an ER graph above, and edge degree correlation coefficient $\alpha \in (-1, +1)$. It has nodes $\mathcal{V} = [n]$, and the random edges $\mathcal{E}$ have degrees (X, Y) with mean correlation $\text{Corr}(\mathsf{X}, \mathsf{Y}) = \alpha$.*

The AER graph $G_\alpha = G_{\alpha,k}$ is obtained from an initial non-assortative ER graph $G_{\epsilon,0} = (\mathcal{V}, \mathcal{E}_0)$ with assortativity α_0 by applying the primitive rewiring operation in Def. 10 k times in succession, each time selecting a pair of edges to rewire, yielding a sequence of graphs $(G_{\epsilon,0}, G_{\alpha,1}, \ldots, G_{\alpha,k})$ with $G_{\alpha,j} = (\mathcal{V}, \mathcal{E}_j)$ and either $\mathcal{E}_j = r_1(\mathcal{E}_{j-1})$ or $\mathcal{E}_j = r_2(\mathcal{E}_{j-1})$ for $j \in [k]^+$.

The assortativity parameter $\alpha = \alpha_k$ for the final AER graph G_α is obtained by repeatedly applying Lem. 7, yielding

$$\alpha_k = \alpha_0 + \sum_{j=1}^{k} \Delta_{j,i(j)}, \tag{4.12}$$

via a Stochastic Rewiring Algorithm (SRA) Chaps. 2 and 3 where, as discussed above, $\alpha_0 \approx 0$, and $\Delta_{j,i(j)}$ is the change in correlation induced by rewiring the edge set from $\mathcal{E}_{j-1}$ to $\mathcal{E}_j$, where $i(j) \in \{1, 2\}$ indicates which of the two rewirings in Def. 10 is selected in stage j.[2]

4.3.2 Joint degree distribution of an AER graph edge

Lemma 8. *If $(\mathsf{X}_n, \mathsf{Y}_n)$ are the random degrees for a randomly selected edge from the random AER graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ with parameters (n, s, α), then X_n and Y_n each have marginal binomial distribution $1 + \mathrm{Bin}(n-2, s)$, with mean $1 + (n-2)s$, variance $(n-2)(1-s)s$, and correlation α.*

$$\begin{aligned} \boldsymbol{\kappa}_n &= (1 + (n-2)s) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ \mathbf{C}_n &= (n-2)s(1-s) \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \end{aligned} \tag{4.13}$$

Proof. Follows from X_n and Y_n being the degrees of the endpoints of a randomly selected edge which are binomial random variables in an AER graph. □

Proposition 8. *Given a bivariate binomial RV $\mathbf{W}_n = [\mathsf{A}_n, \mathsf{B}_n]^\mathsf{T}$ with mean $\boldsymbol{\kappa}_n$ and covariance matrix $\boldsymbol{C}_n$ and assuming $-1 < \alpha < 1$,*

$$\boldsymbol{\kappa}_n = \begin{bmatrix} 1 + (n-2)s \\ 1 + (n-2)s \end{bmatrix}, \ \boldsymbol{C}_n = (n-2)(1-s)s \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \tag{4.14}$$

The standardization of $\mathbf{W}_n$ with zero mean and identity covariance is

$$\tilde{\mathbf{W}}_n = \mathbf{C}_n^{-1/2}(\mathbf{W}_n - \boldsymbol{\kappa}_n) \tag{4.15}$$

See proof in Sec. 4.7.

The following multivariate de Moivre Laplace theorem Veeh [1986], gives conditions under which

[2] It is certainly not the case that a graph *exists* for every choice of $(\mathbf{d}, \alpha)$, and it is an interesting and appears to be an open question to identify the set $\mathcal{A}(\mathbf{d}) \subseteq [-1, +1]$ such that a graph with $|\mathcal{V}| = n$ exists for each $\alpha \in \mathcal{A}(\mathbf{d})$. This question is not the focus of this chapter, however, and is left for future work.

the multivariate binomial distribution converges to the multivariate normal distribution.

Theorem 3 (Multivariate de Moivre Laplace Veeh [1986])**.** *Let* $\{\mathbf{W}_n\}$ *be a sequence of* k*-dimensional binomial random vectors where* $\mathbf{W}_n$ *has mean* $\boldsymbol{\kappa}_n$ *and covariance matrix* $\mathbf{C}_n$*. Then* $\tilde{\mathbf{W}}_n = \mathbf{C}_n^{-1/2}(\mathbf{W}_n - \boldsymbol{\kappa}_n)$ *converges in distribution to the standard* k*-variate normal random vector* $\tilde{\mathbf{Z}}$ *iff* $\|\mathbf{C}_n^{-1}\| \to 0$ *as* $n \to \infty$*.*

Observe that Prop. 8.

$$\mathbf{C}_n^{-1} = \frac{1}{(1+(n-2)s)(1-\alpha^2)} \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} \tag{4.16}$$

and consequently $\|\mathbf{C}_n^{-1}\| \to 0$ as $n \to \infty$, for any matrix norm, as required by the theorem. Thm. 3 and Eq. (4.16) give for the appropriately standardized two element binomial random vector with zero mean and identity covariance

$$\tilde{\mathbf{W}}_n = (\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n)^{\mathsf{T}} = \mathbf{C}_n^{-1/2}(\mathbf{W}_n - \boldsymbol{\kappa}_n) \tag{4.17}$$

converges in distribution to the two element standard normal random vector $(\tilde{\mathsf{Q}}, \tilde{\mathsf{R}})$, $(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) \xrightarrow{D} (\tilde{\mathsf{Q}}, \tilde{\mathsf{R}})$, as $n \to \infty$.

Given a normal random variable $\mathbf{Z}_n$ with elements parameterized by mean $\mu = 1 + (n-2)s$, variance $\sigma^2 = (n-2)s(1-s)$, and covariance α, Prop. 9 states that its normalization is identical to the normalization of binomial random variable $\mathbf{W}_n$.

Proposition 9. *Let* $\mathbf{Z}_n = [\mathsf{Q}_n,\ \mathsf{R}_n]^{\mathsf{T}}$ *be a bivariate normal RV with mean, covariance matrix*

$$\mathbf{m} = \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \qquad \mathbf{D} = \sigma^2 \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \ \textit{where} -1 < \alpha < 1 \tag{4.18}$$

Given the De Moivre Laplace approximation $\mu = 1 + (n-2)s$*,* $\sigma^2 = (n-2)s(1-s)$ *and the corresponding bivariate standard normal RV* $\tilde{\mathbf{Z}} = \mathbf{D}^{-1/2}(\mathbf{Z}_n - \mathbf{m}) = [\tilde{\mathsf{Q}},\ \tilde{\mathsf{R}}]^{\mathsf{T}}$*, where the standardization functions of the elements of* $\mathbf{Z}_n$ *are the same as for those of the elements of* $\mathbf{W}_n$*, i.e.*

$$h_{\tilde{\mathsf{A}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = h_{\tilde{\mathsf{Q}}}(\tilde{\mathsf{Q}}, \tilde{\mathsf{R}}) \quad \textit{and} \quad h_{\tilde{\mathsf{B}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = h_{\tilde{\mathsf{R}}}(\tilde{\mathsf{Q}}, \tilde{\mathsf{R}}) \tag{4.19}$$

Proof. It can be shown that the square root of the inverse of $\mathbf{D}$ is

$$\mathbf{D}^{-1/2} = \frac{1}{2}\begin{bmatrix} x+y & x-y \\ x-y & x+y \end{bmatrix} \text{ where } x = \frac{1}{\sqrt{(1+\alpha)\sigma^2}} \text{ and } y = \frac{1}{\sqrt{(1-\alpha)\sigma^2}} \tag{4.20}$$

Therefore the elements of $\tilde{\mathbf{Z}}$ are $\tilde{\mathsf{Q}} = \frac{1}{2}((x+y)(\mathsf{Q}_n-\mu)+(x-y)(\mathsf{R}_n-\mu))$ and $\tilde{\mathsf{R}} = \frac{1}{2}((x-y)(\mathsf{Q}_n-\mu)+(x+y)(\mathsf{R}_n-\mu))$ and substituting $\sigma^2 = (n-2)s(1-s)$ into x and y it follows from the proof of *Prop.* 8 that $\mathbf{D}^{-1/2} = \mathbf{C}^{-1/2}$ since $x = c$ and $y = b$. This implies that the standardization functions of the elements of $\mathbf{Z}_n$ are the same as the standardization functions for the elements of $\mathbf{W}_n$, i.e. from Eq. (4.68) and Eq. (4.69) $h_{\tilde{\mathsf{A}}_n}(\mathsf{A}_n, \mathsf{B}_n) = h_{\tilde{\mathsf{Q}}}(\mathsf{Q}_n, \mathsf{R}_n)$ and $h_{\tilde{\mathsf{B}}_n}(\mathsf{A}_n, \mathsf{B}_n) = h_{\tilde{\mathsf{R}}}(\mathsf{Q}_n, \mathsf{R}_n)$. □

The inverses of the standardization functions $h_{\tilde{\mathsf{A}}_n}(\mathsf{A}_n, \mathsf{B}_n)$ and $h_{\tilde{\mathsf{B}}_n}(\mathsf{A}_n, \mathsf{B}_n)$ for $\mathbf{W}_n = [\mathsf{A}_n, \mathsf{B}_n]^T$ are given explicitly in Lem. 9.

Lemma 9. *Given the standardized binomial RV* $\tilde{\mathbf{W}}_{\boldsymbol{n}} = [\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n]^{\mathsf{T}}$ *of bivariate binomial RV* $\mathbf{W} = [\mathsf{A}_n, \mathsf{B}_n]^{\mathsf{T}}$ *mean and covariance matrix are* $\boldsymbol{\kappa}_n$ *and* $\boldsymbol{C}_n$*. The inverse of the functions standardizing elements* A_n *and* B_n *are*

$$\mathsf{A}_n = h^{-1}_{\tilde{\mathsf{A}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = \frac{1}{2bc}\left[\tilde{\mathsf{A}}_n(b+c) - \tilde{\mathsf{B}}_n(b-c)\right] + \mu \tag{4.21}$$

$$\mathsf{B}_n = h^{-1}_{\tilde{\mathsf{B}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = \frac{1}{2bc}\left[\tilde{\mathsf{B}}_n(b+c) - \tilde{\mathsf{A}}_n(c-b)\right] + \mu \tag{4.22}$$

where $\mu = (1+(n-2)s)$*,* $b = \frac{1}{\sqrt{(n-2)(s-1)s(\alpha-1)}}$*, and* $c = \frac{1}{\sqrt{(2-n)(s-1)s(\alpha+1)}}$*.*

Proof. Letting $\mu = (1+(n-2)s)$ and solving for B_n in Eq. (4.68) and Eq. (4.69) gives,

$$\mathsf{B}_n = \frac{2\tilde{A}_n - (b+c)(\mathsf{A}_n-\mu)}{c-b} + \mu, \qquad \text{and} \qquad \mathsf{B}_n = \frac{2\tilde{\mathsf{B}}_n - (c-b)(\mathsf{A}_n-\mu)}{b+c} + \mu \tag{4.23}$$

Setting these equations equal to one another and solving for A_n gives Eq. (4.21). The proof of the inverse of the function standardizing B_n is nearly identical. Solving for A_n in Eq. (4.68) and Eq. (4.69) gives,

$$\mathsf{A}_n = \frac{2\tilde{\mathsf{A}}_n - (c-b)(\mathsf{B}_n-\mu)}{b+c} + \mu, \qquad \text{and} \qquad \mathsf{A}_n = \frac{2\tilde{\mathsf{B}}_n - (b+c)(\mathsf{B}_n-\mu)}{c-b} + \mu \tag{4.24}$$

Setting these equations equal to one another and solving for B_n gives Eq. (4.21). □

Corollary 2. *The inverses of the functions standardizing the elements of* $\mathbf{W}_n$ *and* $\mathbf{Z}_n$ *are identical*

i.e.,

$$h^{-1}_{\tilde{\mathsf{A}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = h^{-1}_{\tilde{\mathsf{Q}}}(\tilde{\mathsf{Q}}, \tilde{\mathsf{R}}) \quad \textit{and} \quad h^{-1}_{\tilde{\mathsf{B}}_n}(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) = h^{-1}_{\tilde{\mathsf{R}}}(\tilde{\mathsf{Q}}, \tilde{\mathsf{R}}) \tag{4.25}$$

Proof. This corollary follows directly from Prop. 9. □

The Berry-Esseen bound places an upper bound on the error in convergence of the standard binomial RV $\tilde{\mathbf{W}}$ to the standard normal RV $\tilde{\mathbf{Z}}$. To achieve this Bentkus [2003] introduces the following framework. First let $\mathscr{A}$ be a class of measurable subsets $\mathcal{A} \subset \mathbb{R}^d$ and define ϵ-neighborhoods of sets $\mathcal{A}$ as

Definition 12. *Define ϵ-neighborhoods $\mathcal{A}^\epsilon$ and $\mathcal{A}^{-\epsilon}$ as,*

$$\mathcal{A}^\epsilon = \{x \in \mathbb{R}^d : \rho_{\mathcal{A}}(x) \leq \epsilon\} \quad \textit{and} \quad \mathcal{A}^{-\epsilon} = \{x \in \mathcal{A} : \mathcal{B}_\epsilon(x) \subset \mathcal{A}\} \tag{4.26}$$

where $\rho_{\mathcal{A}}(x) = \inf_{y \in \mathcal{A}} |x - y|$ for and $\mathcal{B}_\epsilon(x) = \{y \in \mathbb{R}^d : |x - y| \leq \epsilon\}$ for $\mathcal{A} \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$.

Second define class $\mathscr{A}$ such that it satisfies the conditions: I) $\mathscr{A}$ is invariant under rescaling: $a\mathcal{A} \in \mathscr{A}$, if $\mathcal{A} \in \mathscr{A}$ for $a \in \mathbb{R}$, $a > 0$. II) $\mathscr{A}$ is shift invariant: $x + \mathcal{A} \in \mathscr{A}$ if $\mathcal{A} \in \mathscr{A}$ and $x \in \mathbb{R}^d$. III) $\mathscr{A}$ is invariant under taking of ϵ-neighborhoods: $\mathcal{A}^\epsilon \in \mathscr{A}$ and $\mathcal{A}^{-\epsilon} \in \mathscr{A}$ if $\mathcal{A} \in \mathscr{A}$ and $\epsilon > 0$.

Third, letting $\mathbb{P}\{\tilde{\mathbf{Z}} \in \mathcal{A}\}$ be a multivariate standard normal distribution for $\mathcal{A} \in \mathscr{A}$, there exist constants $a_d = a_d(\mathscr{A})$ such that

$$\mathbb{P}\{\tilde{\mathbf{Z}} \in \mathcal{A}^\epsilon \backslash \mathcal{A}\} \leq a_d \epsilon, \ \mathbb{P}\{\tilde{\mathbf{Z}} \in \mathcal{A} \backslash \mathcal{A}^{-\epsilon}\} \leq a_d \epsilon \text{ for all } \mathcal{A} \in \mathscr{A} \text{ and } \epsilon > 0 \tag{4.27}$$

Given this framework Bentkus introduces an upper bound on the error in the multivariate normal approximation of a multivariate binomial.

Theorem 4. *(Multivariate Berry-Essen Bentkus [2003]) If class $\mathscr{A}$ satisfies the conditions above, the standard normal distribution $\mathbb{P}\{\tilde{\mathbf{Z}} \in A\}$ satisfies equation Eq.* (4.27)*, and $\tilde{\mathbf{W}}_n$ is the sum of n iid random vectors $\mathbf{B}_0, \ldots, \mathbf{B}_{n-1}$ with, zero mean, identity covariance, and $\bar{\beta} = \mathbb{E}[|\mathbf{B}_i|^3]$ for $i \in [n]$. Then*

$$\Delta_n \leq \frac{100 b_d \bar{\beta}}{\sqrt{n}}, \qquad b_d = \max\{1, a_d\}$$

where

$$\Delta_n \equiv \Delta_n(\mathscr{A}) = \sup_{A \in \mathscr{A}} |\mathbb{P}\{\tilde{\mathbf{W}}_n \in A\} - \mathbb{P}\{\tilde{\mathbf{Z}} \in A\}|$$

and d is the dimension of $\mathbf{B}_i$ for $i \in [n]$.

Letting each $\mathbf{B}_i$ be a standardized Bernoulli RV such that the sum $\tilde{\mathbf{W}}_n = \sum_{i=0}^{n-1} \mathsf{B}_i$ is a standardized binomial RV, Thm. 4 gives an upper bound on the error of the multivariate De Moivre Laplace approximation.

Def. 13 defines asymptotic equivalence between two functions.

Definition 13. *(Breitung [2006]) If for two functions $f : \mathcal{M} \to \mathbb{R}$ and $g : \mathcal{M} \to \mathbb{R}$ (both non-vanishing in a neighborhood of x_0) as $x \to x_0$ with $x \in \mathcal{M}$*

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 1 \tag{4.28}$$

$f(x)$ and $g(x)$ are asymptotically equivalent denoted as $f(x) \sim g(x)$ for $x \to x_0$.

Thm. 5 states the conditions asymptotic equivalence holds between functions $h(f(x))$ and $h(g(x))$ if $f(x)$ and $g(x)$ are asymptoticaly equivalent.

Theorem 5. *(Breitung [2006]) Let there be given two functions $f(x)$ and $g(x)$ on set $\mathcal{M}$ with $f(x) \sim g(x),\ x \to x_0$. Further let there be given a function h on a set $\mathcal{D}$ such that $f(x) \in \mathcal{D}$ and get $g(x) \in \mathcal{D}$ for all $x \in \mathcal{M}$. If there is a closed set $\mathcal{X} \subset \mathcal{D}$ (which may contain infinity) with $f(x) \in \mathcal{X}$ and $g(x) \in \mathcal{X}$ for all $x \in \mathcal{M}$ such that h is continuous on $\mathcal{X}$ and $h(x) \neq 0$ for all $x \in \mathcal{X}$, then $h(f(x)) \sim h(g(x)),\ x \to x_0$.*

Given the Berry Essen bound in Thm. 4 and Def. 13 it follows that the multivariate standard normal distribution and the standard binomial distribution are asymptotically equivalent.

Lemma 10. *If $\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\} > 0$ all sets $\mathcal{A} \in \mathscr{A}$ and for all $n' \in \mathbb{N}^+$ where $\mathbb{N}^+$ is the set of positive integers,*

$$\lim_{n' \to \infty} \frac{\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\}}{\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} \to 1 \tag{4.29}$$

implying $\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\}$ and $\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}$ are asymptotically equivalent.

Proof. From the Berry-Essen bound in Thm. 4 it follows that

$$\sup_{\mathcal{A} \in \mathscr{A}} |\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\} - \mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}| \leq \frac{100 b_d \bar{\beta}}{\sqrt{n'}} \tag{4.30}$$

$$\Rightarrow \quad \frac{1}{\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} \sup_{\mathcal{A} \in \mathscr{A}} |\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\} - \mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}| \leq \frac{100 b_d \bar{\beta}}{\sqrt{n} \mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} \tag{4.31}$$

$$\Rightarrow \quad \sup_{\mathcal{A} \in \mathscr{A}} \left| \frac{\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\}}{\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} - 1 \right| \leq \frac{100 b_d \bar{\beta}}{\sqrt{n} \mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} \tag{4.32}$$

and since $\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\} > 0$ thier exists an arbitrarily small $\epsilon > 0$, $\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\} > \epsilon$ and as $n' \to \infty$,

$$\frac{100 b_d \bar{\beta}}{\sqrt{n'}\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} \to 0 \quad \Rightarrow \quad \sup_{\mathcal{A} \in \mathscr{A}} \left| \frac{\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\}}{\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} - 1 \right| = 0 \quad \Rightarrow \quad \frac{\mathbb{P}\{\tilde{\mathbf{W}}_{n'} \in \mathcal{A}\}}{\mathbb{P}\{\tilde{\mathbf{Z}}_{n'} \in \mathcal{A}\}} = 1 \tag{4.33}$$

□

The proof for Thm. 6 below requires a standard result for the bivariate normal distribution giving the expectation and variance of one of the component random variables conditioned on the value of the other.

Lemma 11 (Chatfield and Collins [1980] §2.3)**.** *Let* $(\mathsf{Q}_n, \mathsf{R}_n)$ *be bivariate normal random variable with:*

$$(\mathsf{Q}_n, \mathsf{R}_n) \sim \mathcal{N}\left(\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \gamma\sigma_1\sigma_2 \\ \gamma\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right). \tag{4.34}$$

Conditioned on $\mathsf{R}_n = j$, Q_n *is normally distributed, i.e.,* $\mathsf{Q}_n | (\mathsf{R}_n = j) \sim \mathcal{N}(\eta_1(j), \sigma_1^2(j))$, *where*

$$\begin{aligned} \eta_1(j) &\equiv \mathbb{E}[\mathsf{Q}_n | \mathsf{R}_n = j] = \eta_1 + \gamma \frac{\sigma_1}{\sigma_2}(j - \eta_2) \\ \sigma_1^2(j) &\equiv \mathrm{Var}(\mathsf{Q}_n | \mathsf{R}_n = j) = \sigma_1^2(1 - \gamma^2). \end{aligned} \tag{4.35}$$

In addition if $\eta_1 = \eta_2 = \mu = 1 + (n-2)s$, $\sigma_1^2 = \sigma_2^2 = \sigma^2 = (n-2)s(1-s)$, *and* $\gamma = \alpha$,

$$\begin{aligned} \mathbb{E}[\mathsf{Q}_n | \mathsf{R}_n = j] &= \alpha j + (1-\alpha)(1 + (n-2)s) \\ \mathrm{Var}(\mathsf{Q}_n | \mathsf{R}_n = j) &= (ns + s^2 - 2s)(1 - \alpha^2). \end{aligned} \tag{4.36}$$

Theorem 6. *Given a bivariate binomial RV* $\mathbf{W}_{n'} = [\mathsf{A}_{n'}, \mathsf{B}_{n'}]^\mathsf{T}$ *and bivariate normal RV* $\mathbf{Z}_{n'} = [\mathsf{Q}_{n'}, \mathsf{R}_{n'}]^\mathsf{T}$ *where* $n' = n-2$ *and* $\mathsf{Q}_{n'}$, $\mathsf{R}_{n'}$ *have mean* $\mu = 1 + (n-2)s$, *variance* $\sigma^2 = (n-2)s(1-s)$, *and correlation coefficient* α *on a class of closed sets* $\mathscr{A}$ *satisfying the conditions for Thm. 4 to hold. By definition* $\mathbb{P}(\mathsf{Z}_{n'} \in \mathcal{A}) > 0$ *for* $\mathcal{A} \in \mathbb{R}^2$ *and therefore as* $n' \to \infty$,

$$\frac{\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]}{\mathbb{E}[\mathsf{Q}_{n'} | \mathsf{R}_{n'} = j]} = \frac{\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]}{\alpha j + (1-\alpha)(1 + (n-2)s)} \to 1 \tag{4.37}$$

and therefore as $n' \to \infty$, $\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]$ *is asymptotically equivalent to* $\alpha j + (1-\alpha)(1 + (n-2)s)$ *denoted* $\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j] \sim \alpha j + (1-\alpha)(1 + (n-2)s)$.

Proof. Notice defining,

$$\mathbb{E}[g_{j,\mu,\sigma}(\tilde{x},\tilde{y})] = \int_{-\infty}^{\infty} \left(\frac{h_{\tilde{\mathsf{X}}}^{-1}(\tilde{x},\tilde{y})}{f_{\mathsf{Y}_{n'}}(j)}\right) \left(\frac{f_{\mathsf{X}_{n'},\mathsf{Y}_{n'}}\left(h_{\tilde{\mathsf{X}}}^{-1}(\tilde{x},\tilde{y}), h_{\tilde{\mathsf{Y}}}^{-1}(\tilde{x},\tilde{y})\right)}{f_{\tilde{\mathsf{X}},\tilde{\mathsf{Y}}}(\tilde{x},\tilde{y})}\right) \mathbb{1}\left(\tilde{y} = k(\tilde{x},j)\right) d_{\tilde{x}} \tag{4.38}$$

where $z(\cdot,\cdot)$ is derived by rearranging Lem. 9 to give $\tilde{\mathsf{B}}_{n'}$ as a function of $(\tilde{\mathsf{A}}_{n'}, \mathsf{B}_{n'})$,

$$\tilde{\mathsf{B}}_{n'} = z(\tilde{\mathsf{A}}_{n'}, \mathsf{B}_{n'}) = \tilde{\mathsf{A}}_{n'} \frac{2bc(\mathsf{B}_{n'} + \mu)}{b + c}, \tag{4.39}$$

$h_{\tilde{\mathsf{X}}}^{-1}(\tilde{x},\tilde{y})$ and $h_{\tilde{\mathsf{Y}}}^{-1}(\tilde{x},\tilde{y})$ are defined in Lem. 9, $f_{\mathsf{Y}_{n'}}(\cdot)$ and $f_{\mathsf{X}_{n'},\mathsf{Y}_{n'}}(\cdot,\cdot)$ are the distribution and joint distribution of an unstandardised RV's $\mathsf{X}_{n'}$, $\mathsf{Y}_{n'}$ which depends on n', and $f_{\tilde{\mathsf{X}},\tilde{\mathsf{Y}}}(\cdot,\cdot)$ is the joint distribution of the standardization of RV's $\mathsf{X}_{n'}$, $\mathsf{Y}_{n'}$ denoted as $\tilde{\mathsf{X}}$, $\tilde{\mathsf{Y}}$. Notice that the standardization of $\mathsf{X}_{n'}$ and $\mathsf{Y}_{n'}$, functions $h_{\tilde{\mathsf{X}}}^{-1}(\tilde{x},\tilde{y})$, $h_{\tilde{\mathsf{Y}}}^{-1}(\tilde{x},\tilde{y})$, and $z(\tilde{\mathrm{x}},j)$ all depend implicitly on μ, σ^2 for $\mathsf{X}_{n'}$ and $\mathsf{Y}_{n'}$ which in turn depend on n', s, and α.

Since $\mathbb{E}[g_{j,\mu,\sigma}(\tilde{x},\tilde{y})]$ satisfies the conditions placed on $h(x)$ in Thm. 5 it follows from Lem. 10 that

$$\lim_{n'\to\infty} \frac{\mathbb{E}[g_{j,\mu,\sigma}(\tilde{\mathsf{A}}_{n'}, \tilde{\mathsf{B}}_{n'})]}{\mathbb{E}[g_{j,\mu,\sigma}(\tilde{\mathsf{Q}}_{n'}, \tilde{\mathsf{R}}_{n'})]} \to 1 \tag{4.40}$$

For notational convience let $\mathbb{E}[g_{j,\mu,\sigma}(\tilde{\mathsf{A}}_{n'}, \tilde{\mathsf{B}}_{n'})] = \bar{g}(\tilde{\mathsf{A}}_{n'}, \tilde{\mathsf{B}}_{n'})$ and notice that

$$\begin{aligned}
\bar{g}(\tilde{\mathsf{A}}_{n'}, \tilde{\mathsf{B}}_{n'}) &= \int_{-\infty}^{\infty} f_{\tilde{\mathsf{A}}_{n'},\tilde{\mathsf{B}}_{n'}}(\tilde{a},\tilde{b}) \left(\frac{h_{\tilde{\mathsf{A}}_{n'}}^{-1}(\tilde{a},\tilde{b})}{f_{\mathsf{B}_{n'}}(j)}\right) \left(\frac{f_{\mathsf{A}_{n'},\mathsf{B}_{n'}}(h_{\tilde{\mathsf{A}}_{n'}}^{-1}(\tilde{a},\tilde{b}), h_{\tilde{\mathsf{B}}_{n'}}^{-1}(\tilde{a},\tilde{b}))}{f_{\tilde{\mathsf{A}}_{n'},\tilde{\mathsf{B}}_{n'}(\tilde{a},\tilde{b})}}\right) \mathbb{1}\left(\tilde{\mathsf{B}}_{n'} = z(\tilde{\mathsf{A}}_{n'}, j)\right) \, d_{\tilde{a}} \, d_{\tilde{b}} \\
&= \int_{-\infty}^{\infty} \left(\frac{h_{\tilde{\mathsf{A}}_{n'}}^{-1}(\tilde{a},\tilde{b})}{f_{\mathsf{B}_{n'}}(j)}\right) \left(f_{\mathsf{A}_{n'},\mathsf{B}_{n'}}(h_{\tilde{\mathsf{A}}_{n'}}^{-1}(\tilde{a},\tilde{b}), h_{\tilde{\mathsf{B}}_{n'}}^{-1}(\tilde{a},\tilde{b}))\right) \mathbb{1}\left(\tilde{\mathsf{B}}_{n'} = z(\tilde{\mathsf{A}}_{n'}, j)\right) \, d_{\tilde{a}} \, d_{\tilde{b}} \qquad (4.41) \\
&= \int_{-\infty}^{\infty} \frac{a}{f_{\mathsf{B}_{n'}}(j)} f_{\mathsf{A}_{n'},\mathsf{B}_{n'}}(a, b = j) \, d_a \qquad (4.42) \\
&= \mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j] \qquad (4.43)
\end{aligned}$$

and additionally

$$\begin{aligned}
\bar{g}(\tilde{\mathsf{Q}}_{n'}, \tilde{\mathsf{R}}_{n'}) &= \int_{-\infty}^{\infty} f_{\tilde{\mathsf{Q}},\tilde{\mathsf{R}}}(\tilde{q},\tilde{r}) \left(\frac{h_{\tilde{\mathsf{Q}}}^{-1}(\tilde{q},\tilde{r})}{f_{\mathsf{R}_{n'}}(j)}\right) \left(\frac{f_{\mathsf{Q}_{n'},\mathsf{R}_{n'}}(h_{\tilde{\mathsf{Q}}}^{-1}(\tilde{q},\tilde{r}), h_{\tilde{\mathsf{R}}}^{-1}(\tilde{q},\tilde{r}))}{f_{\tilde{\mathsf{Q}},\tilde{\mathsf{R}}(\tilde{q},\tilde{r})}}\right) \mathbb{1}\left(\tilde{\mathsf{R}}_{n'} = z(\tilde{\mathsf{Q}}_{n'}, j)\right)\, d_{\tilde{q}}\, d_{\tilde{r}} \\
&= \int_{-\infty}^{\infty} \left(\frac{h_{\tilde{\mathsf{Q}}}^{-1}(\tilde{q},\tilde{r})}{f_{\mathsf{R}_{n'}}(j)}\right) f_{\mathsf{Q}_{n'},\mathsf{R}_{n'}}(h_{\tilde{\mathsf{Q}}}^{-1}(\tilde{q},\tilde{r}), h_{\tilde{\mathsf{R}}}^{-1}(\tilde{q},\tilde{r})) \mathbb{1}\left(\tilde{\mathsf{R}}_{n'} = z(\tilde{\mathsf{Q}}_{n'}, j)\right)\, d_{\tilde{q}}\, d_{\tilde{r}} \qquad (4.44) \\
&= \int_{-\infty}^{\infty} \left(\frac{q}{f_{\mathsf{R}_{n'}}(j)}\right) f_{\mathsf{Q}_{n'},\mathsf{R}_{n'}}(q, r=j)\, d_q \qquad (4.45) \\
&= \mathbb{E}[\mathsf{Q}_{n'} | \mathsf{R}_{n'} = j] \qquad (4.46)
\end{aligned}$$

Therefore given Eq. (4.40), Eq. (4.43), and Eq. (4.46)

$$\begin{aligned}
\lim_{n'\to\infty} \frac{\mathbb{E}[g_{j,\mu,\sigma}(\mathsf{A}_{n'}, \mathsf{B}_{n'})]}{\mathbb{E}[g_{j,\mu,\sigma}(\mathsf{Q}_{n'}, \mathsf{R}_{n'})]} &= \lim_{n'\to\infty} \frac{\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]}{\mathbb{E}[\mathsf{Q}_{n'} | \mathsf{R}_{n'} = j]} \qquad (4.47) \\
&= \lim_{n'\to\infty} \frac{\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]}{\mu + \alpha(j-\mu)} \qquad (4.48) \\
&= \lim_{n'\to\infty} \frac{\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j]}{\alpha j + (1-\alpha)(1+(n-2)s)} \qquad (4.49) \\
&\to 1 \qquad (4.50)
\end{aligned}$$

implying that $\mathbb{E}[\mathsf{A}_{n'} | \mathsf{B}_{n'} = j] \sim \alpha j + (1-\alpha)(1+(n-2)s)$. □

4.3.3 Maximum neighbor degree distribution

This section derives an approximation for the maximum neighbor degree conditional distribution matrix $\mathbf{H}$ for an AER random graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ with parameters (n, s, α) and shows that this approximation converges as $n \to \infty$. Let $\mathsf{Y} = d_{\mathsf{u}}$ be the degree of an NI node u selected uniformly at random from $\bar{\mathcal{V}}$, and let $\mathsf{M} = \lambda_{\mathsf{u}}$ be the maximum neighbor degree of u. The key approximation is that the conditional independence between the degrees of neighboring nodes u conditioned on $\mathsf{Y} = j$ for $j \in \mathcal{D}$.

Lemma 12. *In AER graph G_α if $v \in \Gamma_u$, $\mathsf{Y} = d_u$ and $\mathsf{X}_v = d_v$, then X_v conditioned on Y, $\mathsf{X}_v|\mathsf{Y}$, is independent of the degrees X_q of the nodes $q \in \Gamma_u \backslash \{v\}$.*

Proof. In an ER graph G_ϵ with parameters (n, s), X_v for $v \in \Gamma_u$ is conditionally independent of the degrees X_q of nodes $q \in \Gamma_u \backslash \{u, v\}$ since the edges are placed with uniform probability s. Therefore if X_v is dependent on the degree of a node $q \in \Gamma_u \backslash \{v\}$ in an AER graph G_α, the dependence must be caused by the Stochastic Rewiring Algorithm (SRA) Sec. 4.3.1 when node v was wired to u via edge $e = uv$. But the SRA's wiring of e is independent of the degrees X_q of nodes $q \in \Gamma_u \backslash \{v\}$,

implying $\mathsf{X}_v|\mathsf{Y}$ is independent of X_q for $q \in \Gamma_u \backslash \{v\}$. □

Lemma 13. *Let $G_\alpha = (\mathcal{V}, \mathcal{E})$ be an AER graph with parameters (n, s, α) and maximum degree λ, and recall $\mathbf{H} = (H_{j,k})$ defined in Eq. (4.7). Then $H_{j,k} \to \tilde{H}_{j,k}$ as $n \to \infty$ with*

$$\tilde{H}_{j,k} \equiv \begin{cases} Q_{j,k} - Q_{j,k-1}, & k \in \bar{\mathcal{D}} \setminus \{\lambda\} \\ 1 - Q_{j,\lambda-1}, & k = \lambda \end{cases} \quad \text{where} \quad Q_{j,k} \equiv Q\left(\frac{k - \alpha j - (1-\alpha)(1+(n-2)s)}{\sqrt{(n-2)s(1-s)(1-\alpha^2)}}\right)^j \tag{4.51}$$

and $Q(\cdot)$ is the CDF of the standard normal distribution.

Proof. Of Lem. 13, Fix a random NI node $\mathsf{u} \sim \mathrm{Uni}(\bar{\mathcal{V}})$ with degree $\mathsf{Y} = d_{\mathsf{u}}$. Conditioned on $\mathsf{Y} = j$, let $(\mathsf{X}_v, v \in \Gamma_{\mathsf{u}})$ be the neighbor degrees, with each $\mathsf{X}_v \sim 1 + \mathrm{Bin}(n-2, s)$, and let $\mathsf{M} = \max_{v \in \Gamma_{\mathsf{u}}}(\mathsf{X}_v)$ be the maximum neighbor degree. Therefore,

$$\begin{aligned} H_{j,k} &= \mathbb{P}(\mathsf{M} = k | \mathsf{Y} = j) \\ &= \mathbb{P}(\mathsf{M} \le k | \mathsf{Y} = j) - \mathbb{P}(\mathsf{M} \le k-1 | \mathsf{Y} = j) \\ &= \mathbb{P}(\mathsf{X} \le k | \mathsf{Y} = j)^j - \mathbb{P}(\mathsf{X} \le k-1 | \mathsf{Y} = j)^j \\ &\approx \mathbb{P}(\mathsf{Q} \le k | \mathsf{R} = j)^j - \mathbb{P}(\mathsf{Q} \le k-1 | \mathsf{R} = j)^j \end{aligned} \tag{4.52}$$

where the third equality follows from Lem. 12 and the normal approximation from the edge correlation $\mathrm{Corr}(\mathsf{X}_{\mathsf{v}}, \mathsf{Y})$ being α for random $\mathsf{v} \sim \mathrm{Uni}(\Gamma_u)$ by Def. 9, and Thm. 6 as $n \to \infty$.

Standardizing the conditional normal RV $\mathsf{Q}|\mathsf{R} = j$ gives Eq. (4.51),

$$\begin{aligned} Q_{j,k} &\equiv \mathbb{P}(\mathsf{Q} \le k | \mathsf{R} = j)^j \\ &= \mathbb{P}\left(\frac{\mathsf{Q} - \mathbb{E}[\mathsf{Q}|\mathsf{R}=j]}{\mathrm{Std}(\mathsf{Q}|\mathsf{R}=j)} \le \frac{k - \mathbb{E}[\mathsf{Q}|\mathsf{R}=j]}{\mathrm{Std}(\mathsf{Q}|\mathsf{R}=j)}\right)^j \end{aligned} \tag{4.53}$$

□

4.4 A Markov chain model for max degree search

This section first introduces the essential idea of SAWJ in the form of Alg. 2, Walk-Jump (WJ). Recall the mean absorption time of a discrete-time Markov chain in Sec. 4.4.1. This section creates a Markov chain approximating Alg. 2 for an *arbitrary* graph in terms of its max neighbor degree distribution $\mathbf{H}$ in Sec. 4.4.2, and a Markov chain approximating Alg. 2 for an AER graph in terms of the parameters (n, s, α, λ) in Sec. 4.4.3. The expected fraction of nodes that are local maxima

is estimated for arbitrary and AER graphs in Sec. 4.4.4, and results showing the accuracy of the approximations are in Sec. 4.4.5.

Algorithm 2 Walk-Jump: find any max-degree node $v \in \mathcal{V}_\lambda$

```
require graph G = (V, E), max degree λ = max_{u∈V} d_u
while max{d_u, λ_u} < λ (i.e., V_λ not found) do
  if d_u < λ_u (i.e., have higher degree neighbor) then
    select random v from Γ_u^+ (i.e., walk)
  else (i.e., at local maximum)
    select random v from V (i.e., jump)
  end if
end while
```

4.4.1 Mean absorption time for a discrete-time Markov chain

Let T_U denote the random absorption time of an absorbing discrete-time Markov chain (DTMC) $U \equiv (U(t), t \in \mathbb{N})$ with finite state space $\mathcal{U}$, where $U(0)$ has distribution $\mathbf{l} \equiv (l_u, u \in \mathcal{U})$. Partition $\mathcal{U}$ into transient $\check{\mathcal{U}}$ and absorbing $\hat{\mathcal{U}}$ states where $|\check{\mathcal{U}}| = \check{n}$ and $|\hat{\mathcal{U}}| = \hat{n}$, and partition the transition probability matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{array}{c} \\ \hat{n} \\ \check{n} \end{array} \begin{array}{c} \begin{array}{cc} \hat{n} & \check{n} \end{array} \\ \begin{bmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{M} & \mathbf{L} \end{bmatrix} \end{array}. \tag{4.54}$$

Here $\mathbf{L}, \mathbf{M}, \mathbf{A}$, are submatrices corresponding to transitions between transient states, from transient to absorbing states, and among absorbing states, respectively, and $\mathbf{O}$ is a zeros matrix. The *fundamental matrix*, defined below, is the key to the mean absorption time, as shown in Thm. 7.

Definition 14. *The* fundamental matrix *for the DTMC U is $\mathbf{N} = (\mathbf{I} - \mathbf{L})^{-1}$, for $\mathbf{I}$ the identity matrix. The* mean absorption time *starting from state u is*

$$\boldsymbol{\tau}_u \equiv \mathbb{E}[\mathsf{T}_U | U(0) = u] = \mathbb{E}[\min\{t \in \mathbb{N} | U(t) \in \hat{\mathcal{U}}, U(0) = u)]. \tag{4.55}$$

Theorem 7. *(Kemeny and Snell [1983] §3.3) The mean absorption time is $\mathbb{E}[\mathsf{T}_U] = \sum_{u \in \mathcal{U}} \boldsymbol{\tau}_u l_u$, where $\boldsymbol{\tau}_u = (\mathbf{N}\mathbf{1})_u$, and $\mathbf{1}$ is a vector of ones.*

4.4.2 A Markov chain model for an arbitrary graph

This sections uses Thm. 7 to compute the mean time to find a node in $\mathcal{V}_\lambda$ under Alg. 2. Computing this time requires fixing a graph $G = (\mathcal{V}, \mathcal{E})$, with degree set $\mathcal{D}$, max degree λ, maximum degree nodes $\mathcal{V}_\lambda$, degree distribution $\mathbf{w} = (w_k, k \in \mathcal{D})$, and maximum neighbor degree distribution $\mathbf{H}$.

Using the notation in Sec. 4.4.1, define the discrete-time Markov chain $U = (U(t), t \in \mathbb{N})$, where $U(t)$ represents the (random) degree of the node occupied at time t, the state space is $\mathcal{U} = \mathcal{D}$, the absorbing state is $\hat{\mathcal{U}} = \{\lambda\}$, the transient states are $\check{\mathcal{U}} = \mathcal{D} \setminus \{\lambda\}$, and the initial distribution $\mathbf{l}$ is $\mathbf{w}$ the degree distribution of G. The transition submatrix $\mathbf{P}_U$ of DTMC U has entries

$P_{j,k} \equiv \mathbb{P}(U(t+1) = k | U(t) = j)$, where, for $j \in \check{\mathcal{U}}$:

$$P_{j,k} = \begin{cases} w_k, & j = 0 \\ h_j w_k, & j > 0, k \leq j \\ h_j w_k + H_{j,k}, & 0 < j < k < \lambda \end{cases} \tag{4.56}$$

Here, $\mathbf{h} \equiv (h_j, j \in \mathcal{D})$ has $h_j \equiv \sum_{k \leq j} H_{j,k}$ equal to the probability the max neighbor degree of a degree j node is j or lower, i.e., the probability that a degree j node is a local maximum.

Theorem 8. *The mean time to find a maximum degree node for an arbitrary graph G under Alg. 2, starting from a node selected uniformly at random from $\mathcal{V}$, is given by Thm. 7 for the absorbing Markov chain described in Eq.* (4.56)*.*

Proof. This theorem requires justification of the given transition matrix $\mathbf{P}_U$. Consider a transient state $j \in \check{\mathcal{U}}$. There are two types of state transitions: *i*) from j to some $k \in \check{\mathcal{U}}$ with $k > j$ when node u has a maximum neighbor degree of $k = \lambda_u > j$, or *ii*) from j to any value $k \in \mathcal{D}$ when the maximum neighbor degree of u is j or lower. Then, conditioning on M the maximum neighbor degree:

$$\begin{aligned} P_{j,k} \quad = \quad & \mathbb{P}(U(t+1) = k | \mathsf{M} \leq j, U(t) = j) \mathbb{P}(\mathsf{M} \leq j | U(t) = j) + \\ & \mathbb{P}(U(t+1) = k | \mathsf{M} > j, U(t) = j) \mathbb{P}(\mathsf{M} > j | U(t) = j). \end{aligned} \tag{4.57}$$

For any k, the first term simplifies to $w_k h_j$, since the event $\{\mathsf{M} \leq j\}$ corresponds to node j being a local maximum node, which has probability $h_j = \mathbb{P}(\mathsf{M} \leq j | U = j)$, and results in the decision to jump to a random node, which will have degree k with probability w_k. For $k > j$ the second term simplifies to $H_{j,k} = \mathbb{P}(\mathsf{M} = k | U = j)$, while for $k \leq j$ the second term is zero, since the algorithm will never walk to a node of degree $k \leq j$. □

4.4.3 A Markov chain model for an AER random graph

Whereas Thm. 8 applies to an *arbitrary* graph, and requires knowledge of the max neighbor degree distribution $\mathbf{H}$ to compute the transition submatrix $\mathbf{P}_U$, this section considers an AER random

graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ with parameters (n, s, α, λ) if the maximum degree λ is known. If λ is unknown, the maximum degree can be approximated using extreme value theory as the expected maximum of n binomial RV's $\lambda \approx \mu + \sigma\left(\sqrt{2\log n} - \frac{\log(4\pi \log n)}{2\sqrt{2\log n}} + \frac{\gamma}{\sqrt{2\log n}}\right)$ where γ is Euler-Mascheroni constant Chap. 8.

Using the notation in Sec. 4.4.1, define the discrete-time Markov chain $\tilde{U} = (\tilde{U}(t), t \in \mathbb{N})$, where $\tilde{U}(t)$ represents the (random) degree of the node occupied at time t, the state space is $\mathcal{U} = \tilde{\mathcal{D}} = \{0, \ldots, \lambda\}$, the absorbing state is $\hat{\mathcal{U}} = \{\lambda\}$, the transient states are $\check{\mathcal{U}} = \tilde{\mathcal{D}} \setminus \{\lambda\}$, and the initial distribution $\mathbf{l}$ is $\tilde{\mathbf{w}} = (\tilde{w}_j, j \in \tilde{\mathcal{D}})$ with entries

$$\tilde{w}_j = \frac{\mathbb{P}(\mathsf{X} = j)}{\mathbb{P}(\mathsf{X} \leq \lambda)},\ j \in \tilde{\mathcal{D}}, \tag{4.58}$$

for $\mathsf{X} \sim \text{Bin}(n-1, s)$, i.e., $\tilde{w}_j$ is the probability a binomial RV equals j conditioned on the RV taking value at most λ. The transition submatrix $\mathbf{P}_{\tilde{U}}$ has entries $\tilde{P}_{j,k} \equiv \mathbb{P}(\tilde{U}(t+1) = k | \tilde{U}(t) = j)$, where, for $j \in \check{\mathcal{U}}$ (recall Lem. 13):

$$\tilde{P}_{j,k} = \begin{cases} \tilde{w}_k, & j = 0 \\ Q_{j,j}\tilde{w}_k, & j > 0, k \leq j \\ Q_{j,j}\tilde{w}_k + (Q_{j,k} - Q_{j,k-1}), & 0 < j < k < \lambda \end{cases} \tag{4.59}$$

Theorem 9. *The approximate mean time to find a maximum degree node for the above AER random graph G_α under Alg. 2, starting from a node selected uniformly at random from $\mathcal{V}$, is given by Thm. 7 for the Markov chain described in Eq.* (4.59)*. Further the approximate mean time to find a maximum degree node converges to the mean time to find a maximum degree node as $n \to \infty$.*

Proof. The proof of Thm. 9 follows from the proof of Thm. 8 by letting Y denote a random degree and recognizing the approximations $h_j = \mathbb{P}(\mathsf{M} \leq j | \mathsf{Y} = j) \to Q(j,j)$, $H_{j,k} \to (Q_{j,k} - Q_{j,k-1})$ for $k < \lambda$, and $H_{j,\lambda} \to (1 - Q_{j,\lambda-1})$ from Lem. 13 as $n \to \infty$. □

4.4.4 Approximate fraction of local maxima nodes

Alg. 2 jumps at a local maximum node, i.e., a node with $d_u \geq \lambda_u$, and as such the absorption time is sensitive to the fraction of strict $f_{\text{str}} \equiv |\mathcal{V}_{\text{str}}|/\bar{n}$ and non-strict $f_{\text{nst}} \equiv |\mathcal{V}_{\text{nst}}|/\bar{n}$ local maxima, where $\mathcal{V}_{\text{str}} \equiv \{u \in \bar{\mathcal{V}} | d_u > \lambda_u\}$, and $\mathcal{V}_{\text{nst}} \equiv \{u \in \bar{\mathcal{V}} | d_u \geq \lambda_u\}$ are the corresponding sets of local maximum NI nodes. Isolated nodes are not considered maxima. Let f_{str} and f_{nst} be the RVs denoting f_{str} and f_{nst} in a randomly selected graph G or AER graph G_α

Proposition 10. *For a graph $G = (\mathcal{V}, \mathcal{E})$ with conditional maximum neighbor degree distribution* $\mathbf{H}$*:*

$$\mathbb{E}[\mathsf{f}_{\rm str}] = \tilde{f}_{\rm str} = \sum_{j \in \bar{\mathcal{D}}} (h_j - H_{j,j}) \bar{w}_j, \qquad \mathbb{E}[\mathsf{f}_{\rm nst}] = \tilde{f}_{\rm nst} = \sum_{j \in \bar{\mathcal{D}}} h_j \bar{w}_j, \tag{4.60}$$

for $\mathbf{h} = (h_j, j \in \bar{\mathcal{D}})$ with $h_j = \sum_{k \le j} H_{j,k}$, and $\bar{\mathbf{w}} = (w_j, j \in \bar{\mathcal{D}})$ denoting the degree distribution of NI nodes. For an AER graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ with parameters (n, s, α, λ):

$$\mathbb{E}[\mathsf{f}_{\rm str}] \to \hat{f}_{\rm str} = \sum_{j \in [\lambda]} Q_{j,j-1} \tilde{w}_j, \qquad \mathbb{E}[\mathsf{f}_{\rm str}] \to \hat{f}_{\rm nst} = \sum_{j \in [\lambda]} Q_{j,j} \tilde{w}_j, \tag{4.61}$$

with $Q_{j,k}$ in Eq. (4.51)*, $\tilde{\mathbf{w}} = (\tilde{w}_j, j \in \tilde{\mathcal{D}})$ in Eq.* (4.58) *is a truncated binomial distribution, and the convergences in Eq.* (4.61) *follow from Lem. 13 as $n \to \infty$.*

Proof. For a randomly selected graph $G = (\mathcal{V}, \mathcal{E})$ with distribution $\mathbf{H}$, condition on the degree $\mathsf{Y} = d_{\mathsf{u}}$ of a node $\mathsf{u} \sim \text{Uni}(\bar{\mathcal{V}})$, and let $\mathsf{M} = \lambda_{\mathsf{u}}$ be the maximum neighbor degree:

$$\mathbb{E}[\mathsf{f}_{\rm str}] = \sum_{j \in \bar{\mathcal{D}}} \mathbb{P}(\mathsf{M} < j | \mathsf{Y} = j) \mathbb{P}(\mathsf{Y} = j) = \sum_{j \in \bar{\mathcal{D}}} (h_j - H_{j,j}) \bar{w}_j. \tag{4.62}$$

For a randomly selected AER graph $G_\alpha = (\mathcal{V}, \mathcal{E})$ with parameters (n, s, α, λ):

$$\mathbb{E}[\mathsf{f}_{\rm str}] = \sum_{j \in \bar{\mathcal{D}}} \mathbb{P}(\mathsf{M} < j | \mathsf{Y} = j) \mathbb{P}(\mathsf{Y} = j) \to \sum_{j \in [\lambda]} Q_{j,j-1} \tilde{w}_j, \tag{4.63}$$

with the convergence following from Lem. 13 as $n \to \infty$. The derivations for $\mathbb{E}[\mathsf{f}_{\rm nst}]$ are nearly identical replacing $< j$ with $\le j$, $(h_j - H_{j,j})$ with h_j, and $Q_{j,j-1}$ with $Q_{j,j}$. □

4.4.5 Evaluation of the accuracy of the approximations

Fig. 4.2 shows numerical and simulation results illustrating the accuracy of the approximations used in obtaining the mean absorption time in Thm. 8, Thm. 9, and the fraction of locally maximal nodes given in Prop. 10. The graphs have $n = 1000$ nodes and edge probability $s = 0.005$, yielding mean degree $\mu = 5$, α is swept over $[-0.8, +0.8]$, and the results are averaged over 10 AER graphs for each α.

The accuracy of Thm. 8 and Thm. 9 is shown in Fig. 4.2 (left). The red curve labeled WJ is obtained by averaging the absorption time over 1000 trials of Alg. 2 on each graph and α value (for $10,000$ simulations per α). Curves DTMC Model and AER DTMC Model are found by computing the mean absorption time for the Markov chains in Thm. 8 (using matrix $\mathbf{H}$) and Thm. 9 (using

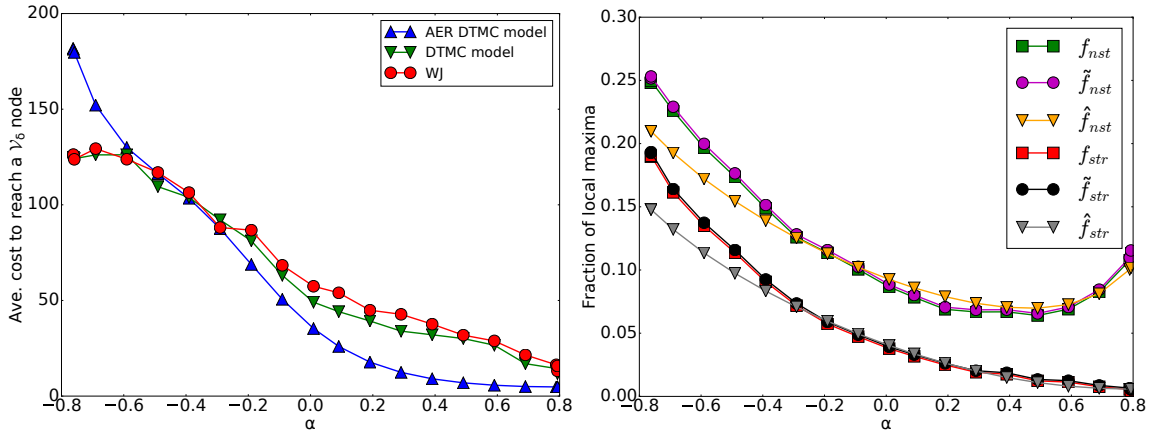

Figure 4.2: Numerical and simulation results to measure the accuracy of the approximations used in the Markov models predicting the mean search time in Thm. 8 and Thm. 9 (left) and the mean fraction of local maxima nodes in Prop. 10 (right), vs. the assortativity α. See Sec. 4.4.5 for explanation.

parameters (n, s, α)), respectively. Note Thm. 8 is *very* accurate, and Thm. 9 is reasonably accurate.[3]

The accuracy of Prop. 10 is shown in the six curves in Fig. 4.2 (right): three curves each for the fraction of strict and non-strict local maxima. For each set, the three curves show the actual fraction, the approximation for the graph using its maximum neighbor degree distribution $\mathbf{H}$ (denoted $\tilde{f}$), and the AER approximation using (n, s, α, λ) (denoted $\hat{f}$). The results show both approximations are fairly accurate for a wide range of α, though $\tilde{f}$ outperforms $\hat{f}$.[4]

4.5 The self-avoiding walk-jump (SAWJ) algorithm

The search time of Alg. 2 was approximated by the Markov chains in Sec. 4.4.2 and Sec. 4.4.3. Sec. 4.5.1 improves Alg. 2 introducing Alg. 3 (SAWJ) and describes several competing algorithms from the literature, Sec. 4.5.2 defines the unit and linear cost models. Next this section looks at the performance (expected cost) of SAWJ against competing algorithms on both AER graphs under the unit and linear cost models (Sec. 4.5.3, Sec. 4.5.4), and on real-world large graphs under the same cost models (Sec. 4.5.5, Sec. 4.5.6).

4.5.1 Algorithm descriptions

The self-avoiding walk-jump (SAWJ) algorithm is given in Alg. 3. It requires a graph $G = (\mathcal{V}, \mathcal{E})$, the maximum degree λ, an algorithm bias parameter $\beta \in [0, 1]$, and assumes $\alpha \geq 0$. The key addition in Alg. 3 relative to Alg. 2 (WJ) is the self avoidance of the walk. If $\alpha < 0$, SAWJ is modified to

[3]The AER DTMC curve uses the *actual* λ for each graph, not an approximation of λ.
[4]The curves $\hat{f}_{nst}$ and $\hat{f}_{str}$ use standard de Moivre Laplace integration corrections.

walk to minimum degree unvisited neighbors under the supposition that $\mathcal{V}_\lambda$ nodes in graphs with negative assortativity have *low* degree neighbors. The SAWJ algorithm thus modified jumps if it is at local minima i.e. $\lambda_u < \lambda$ where λ_u is the maximum degree neighbor of u.

Algorithm 3 Self-avoiding walk–jump (SAWJ, SJ): find a $v \in \mathcal{V}_\lambda$

require graph $G = (\mathcal{V}, \mathcal{E})$, max deg. λ, bias $\beta \in [0, 1]$
initialize $u \in \mathcal{V}$ *(starting node)*, $\mathcal{H} := \emptyset$ *(init. history)*
while $\max\{d_u, \lambda_u\} < \lambda$ *(not yet found $\mathcal{V}_\lambda$)* **do**
  $\mathcal{H} := \mathcal{H} \cup \{u\}$ *(update history)*
  $\lambda_u := \max_{v \in \Gamma_u \setminus \mathcal{H}} d_v$ *(max neighbor degree)*
  $\Gamma_u^+ := \operatorname{argmax}_{v \in \Gamma_u \setminus \mathcal{H}} d_v$ *(max deg. neighbors)*
  if $\Gamma_u^+ \neq \emptyset$ **then** *(there are unvisited neighbors)*
    if $\lambda_u \geq d_u$ *(i.e., not at a local max)* **then**
      select random $v \in \Gamma_u^+$ *(i.e., walk)*
    else *(i.e., at a local max)*
      w.p. β select random $v \in \mathcal{V} \setminus \mathcal{H}$ *(i.e., jump)*
      w.p. $1 - \beta$ select random $v \in \Gamma_u^+$ *(i.e., walk)*
    end if
  else *(no unvisited neighbors)*
    select random $v \in \mathcal{V} \setminus \mathcal{H}$ *(i.e., jump)*
  end if
end while

The remainder of this section sketches the algorithms SAWJ is compared against.

1. Star sampling with replacement (SS-R) (Kolaczyk [2009]) Repeatedly select a random node $u \in \mathcal{V}$ until $d_u = \lambda$ or $\lambda_u = \lambda$.

2. Star sampling without replacement (SS-S) (Kolaczyk [2009]) A history $\mathcal{H}$ is maintained of sampled nodes and their neighbors, and a random node $u \in \mathcal{V} \setminus \mathcal{H}$ is repeatedly selected until $d_u = \lambda$ or $\lambda_u = \lambda$.

3. Albatross sampling (AL) (Jin et al. [2011]) Repeatedly do the following: jump (with probability 0.2) to a random node or walk (w.p. 0.8) to a neighboring node selected with uniform probability from its set of neighbors.

4. Frontier sampling (FS) (Ribeiro and Towsley [2010]) Arbitrarily assuming 5 simultaneous random walkers on the graph; at each time step one of the walkers is selected in proportion to their current nodes degree, and moves to a randomly selected neighboring node.

5. Avrachenkov Walk-Jump (Avra-W, AW) (Avrachenkov et al. [2012]) Setting the termination condition to be identifying a maximum degree node. On each step the walk transitions from $v \in \Gamma_v$

to a neighboring node $u \in \Gamma_v$ selected uniformly at random with probability $p = \left(\frac{\mu}{n} + 1\right)/(d_v + \mu)$ and jumps to a node $u \in \mathcal{V} \backslash \Gamma_v$ with probability $q = 1 - p$ where μ is the graphs expected degree.

6. Avrachenkov Sampling (Avra-S, AS) (Avrachenkov et al. [2014]) Setting the termination condition to identifying a maximum degree node. Each iteration i a node is added to a set of randomly sampled nodes $\mathcal{V}_s$ until $d_u = \lambda$ where $u = \text{argmax}_{v \in \mathcal{V}}(\mathbf{c}_{i_v})$ and $\mathbf{c}_{i_v} = \sum_{q \in \mathcal{V}_s} \mathbb{1}\{v \in \Gamma_q, v \in \mathcal{V}_s\}$.

To compare AL and FS with SAWJ, both AL and FS are terminated if either are in the neighborhood of a node $v \in \mathcal{V}_\lambda$. Analogous modifications where not made to Avra-W and Avra-S as both where designed under the assumption that the degrees of the nodes $u \in \Gamma_v$ are not given on sampling v.

4.5.2 Definition of cost models

How efficient a graph sampling algorithm is depends on the cost of a sample and the information gained per sample. The cost models of graph sampling algorithms are often not explicitly defined Stokes and Weber [2016] or are defined in terms of a particular application such as sampling the Twitter graph Avrachenkov et al. [2014]. Letting v be the sampled node there are four pieces of information that the algorithms consided rely on 1) the index of the sampled node i_v, 2) the degree of the sampled node d_v, 3) the index's of the nodes in v's neighborhood i_u for $u \in \Gamma_v$, and 4) the degrees of the nodes in v's neighborhood d_u for $u \in \Gamma_v$.

Definition 15. *Letting C_v be a RV denoting the cost of sampling node v and $\mathcal{I}$ the information obtained; define two cost models.*

- *Unit cost model:* $\mathsf{C}_v = 1$, $\mathcal{I} \equiv \{i_v, d_v, i_u, d_u | u \in \Gamma_v\})$
- *Linear cost model:* $\mathsf{C}_v = 1$, $\mathcal{I} \equiv \{i_v, d_v, i_u | u \in \Gamma_v\})$

The key difference between these models is to obtain the degree's of v's neighbors under the linear cost model requires sampling each neighbor while this is not necessary under the unit cost model. Therefore the cost of acquiring the same information as the unit cost model under the linear cost model for SS-R, SS-S, SAWJ, and the modified versions of AL and FS is sampled is $|\Gamma_v|$ for $v \in \mathcal{V}$.

4.5.3 Algorithm performance on AER graphs with unit cost model

The average performance of the SAWJ and WJ algorithms from Sec. 4.5.1 and of the competing algorithms under the unit cost model SS-R, SS-S, AL, FS, on AER graphs are shown in Fig. 4.4 and Fig. 4.5 as a function of assortativity α. The graphs used in generating Fig. 4.4's results have

$n = 1000$ nodes and edge density $s = 0.005$, with $(\mu, \lambda) = (5, 13)$ and α is swept over $[-0.8, +0.8]$. The graphs used to generate the results for Fig. 4.5 have $n = 1000$ nodes and edge density $s = 0.02$, with $(\mu, \lambda) = (20, 35)$ and α is swept over the range of achievable assortativity $\approx [-0.4, +0.4]$. For each value of α 10 AER graphs were constructed. Each of the algorithms was run 1000 times on each graph, for $10,000$ trials per α point.

Several points bear mention. I) Avra-S and Avra-W perform significantly worse then the competing algorithms under the unit cost model with the expected cost for Avra-S being within $(1.1 \times 10^4, 1.3 \times 10^7)$ and Avra-W within $(572, 614)$ for $s = 0.005$ while Avra-S within $(6.6 \times 10^3, 7.2 \times 10^3)$ and Avra-W within $(720, 750)$ for $s = 0.02$ respectively. Excluding Avra-S and Avra-W in Fig. 4.4 and Fig. 4.5 it is clear that the AL and FS algorithms are not competitive in either graph sets for most values of α. II) comparing SAWJ vs. WJ, observe the self-avoiding property and following Γ_u^- can halve the search time for $\alpha \approx -1$, at $\alpha \approx 0$ WJ can outperform SAWJ, and for $\alpha > 0$ SAWJ slightly outperforms WJ in both sets of AER graphs. The $\alpha > 0$ case shows self-avoidance is also beneficial, however the benefits may not be pronounced if both SAWJ and WJ revert to random sampling or follow an increasing degree gradient. III) in both Fig. 4.4 and Fig. 4.5 observe SAWJ has a slight upward trend in α for $\alpha < 0$, a "spike" at $\alpha = 0$, and a downward trend for $\alpha > 0$. The improvement as $\alpha \uparrow 1$ ($\alpha \downarrow -1$) reflects the increased value of following maximum (minimum) degree neighbors for increasingly assortative (disassortative) graphs, respectively; the spike at $\alpha = 0$ suggests following a negative degree gradient is not optimal in slightly disassortative graphs. IV) the simulated results suggest the optimal choice of bias parameter β for the SAWJ algorithm depends upon (s, α), although $\beta = 0$ often performs well. V) although SS-S is optimal for low-density ($s = 0.005$) disassortative ($\alpha < 0$) graphs, SAWJ performs better on high-density or assortative graphs, compare Fig. 4.4 and Fig. 4.5.

A surprising observation is that in Fig. 4.4 it can be seen that star sampling without replacement (SS-S) is not always superior to star sampling with replacement (SS-R). Fig. 4.3 gives an example to illustrate how this may occur and Chap. 6 goes into further detail on this phenomena.

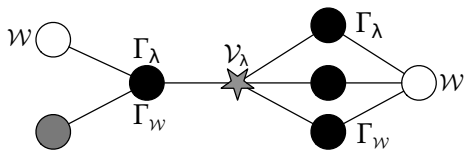

Figure 4.3: Example showing star-sampling without replacement (SS-S) may be inferior to star-sampling with replacement (SS-R). The two white nodes ($\mathcal{W}$) have been sampled; these two nodes and their four neighbors, the black nodes $\Gamma_\mathcal{W}$, have been removed from the sampling pool in SS-S. The maximum degree node is the star ($\mathcal{V}_\lambda$); the black nodes are *also* the neighbors of $\mathcal{V}_\lambda$, denoted Γ_λ. The probability of reaching $\mathcal{V}_\lambda$ or its neighbors Γ_λ on the next sample is 1/2 without replacement and 5/8 with replacement.

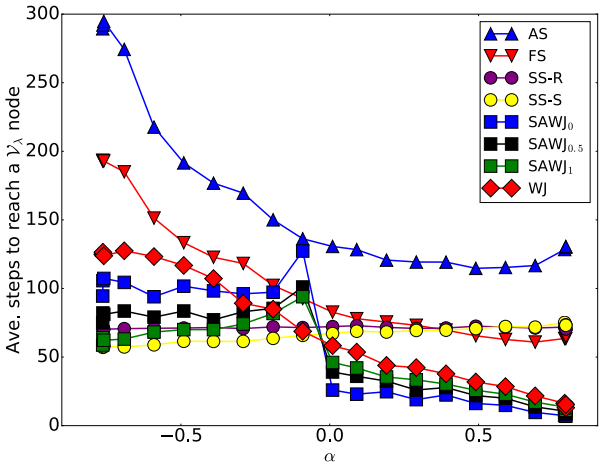

Figure 4.4: $n = 1000$, $s = 0.005$; Unit cost model; (Excluding Avra-S and Avra-W)

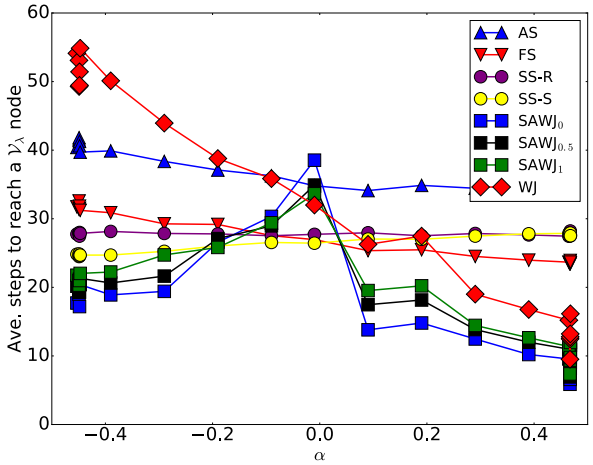

Figure 4.5: $n = 1000$, $s = 0.02$; Unit cost model; (Excluding Avra-S and Avra-W)

4.5.4 Algorithm performance on AER graphs with linear cost model

The average cost to find a maximum degree node for the algorithms in Sec. 4.5.1, excluding Avra-S, on the same AER graphs in Sec. 4.5.3 are shown in Fig. 4.6 and Fig. 4.7 under the linear cost model. I) Using Avra-S costs significantly more to find a node $v \in \mathcal{V}_\lambda$ then the competing algorithms with its expected cost ranging between $(1.1 \times 10^4, 1.3 \times 10^7)$ and $(6.6 \times 10^3, 7.2 \times 10^3)$ for $s = 0.005$ and $s = 0.02$ respectively. II) Neither AS nor FS algorithms are competitive under the linear cost model. III) Comparing SAWJ and WJ, similar results are observed to those of the unit cost model with SAWJ out performing WJ significantly for $\alpha < 0$ and performing comparably for $\alpha \geq 0$. IV) In both Fig. 4.6 and Fig. 4.7 observe SAWJ has an upward trend in α for $\alpha < 0$, a "spike" at $\alpha = 0$, and a downward trend for $\alpha > 0$. V) under the linear cost model although SS-S is optimal for low-density ($s = 0.005$) disassortative ($\alpha < 0$) graphs, SAWJ performs better on high-density or assortative graphs.

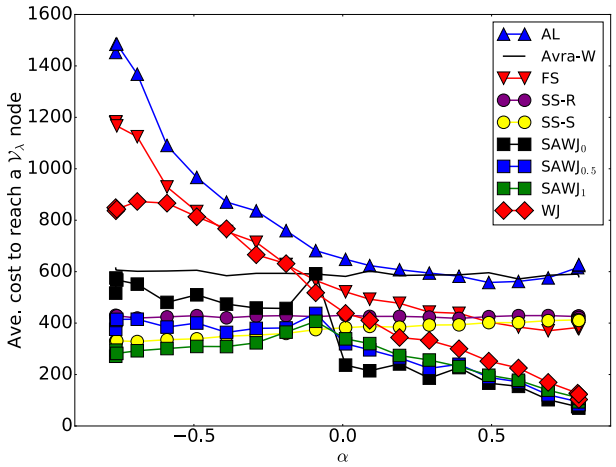

Figure 4.6: $n = 1000$, $s = 0.005$; Linear cost model (Excluding Avra-Sample)

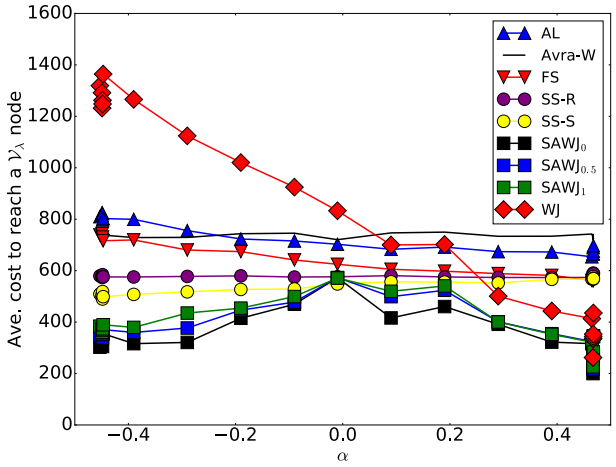

Figure 4.7: $n = 1000$, $s = 0.02$; Linear cost model (Excluding Avra-Sample)

4.5.5 Algorithm performance on real-world graphs with unit cost model

This section tests the SAWJ algorithm on seven real-world graphs GrQc (GrQ), amazon (ama), dblp (dbl), Enron (Enr), EuAll (EuA), facebook (fac), and Gnutella (Gnu) from the Stanford Large Network Dataset Collection (SNAP) Leskovec and Krevl [2014]. Tab. 4.1 gives the statistics for these graphs where n is the number of nodes, m the number of edges, μ the average degree, α the assortativity, λ the maximum degree, $|\mathcal{V}_\lambda|$ the number of maximum degree nodes, ϕ the diameter of the graph, and ▲ is the fraction of triangles in the graph.

Assuming the unit cost model, Tab. 4.2 gives the expected cost required by each algorithm to find a degree λ node averaged over $1,000$ trials per graph and Tab. 4.3 gives the absolute relative error of the expected cost relative to the minimum cost of any of the algorithms. Notice the difference in performance between assortative and disassortative graphs. In GrQc and dblp, both of which are assortative graphs, SAWJ outperforms all competing algorithms and in the facebook graph SAWJ performs comparably to SS-R and SS-S. On the disassortative graphs amazon, Enron, and EuAll the relative performance, $r_p = \frac{steps}{opt}$, of SAWJ compared to the optimal performance of SS-S is $r_p \leq 1.3$ with optimal parameterization $\beta = 1$ implying that SAWJ is essentially star sampling these graphs. In the Gnutella graph FS outperforms the other algorithms and SAWJ's absolute relative error in the cost of employing SAWJ is $r_p = 1.3$. Surprisingly given the statistics of the Gnutella graph are similar to the Enron graph the optimal parameterization of SAWJ on the Gnutella graph is $\beta = 0$, but perhaps SAWJ is less likely to backtrack on the Gnutella graph which is sparse and contains fewer triangles.

Excluding the Gnutella graph in the unit cost model AL, FS, Avra-W, and Avra-S all perform

significantly worse then SS-R, SS-S, or SAWJ.

Table 4.1: Statistics of the SNAP graphs

	n	m	μ	α	λ	$\lvert\mathcal{V}_\lambda\rvert$	ϕ	▲
GrQ	5,242	14,496	5.53	0.66	81	1	17	0.36
ama	334,863	925,872	5.53	-0.06	549	1	44	0.08
dbl	317,080	1,049,866	6.62	0.27	343	1	21	0.13
Enr	36,692	183,831	10.02	-0.11	1.383	1	11	0.03
EuA	265,214	365,570	2.76	-0.18	7,636	1	14	0.001
fac	4,039	88,234	43.69	0.06	1045	1	8	0.26
Gnu	62,586	147,892	4.73	-0.09	95	1	11	0.001

Table 4.2: Expected cost in the unit cost model

	AL	FS	SS-S	SS-R	SJ_0	$SJ_{0.5}$	SJ_1	Avra-W	Avra-S
GrQ	550	150	64	61	11	**8**	9	734	5,725
ama	6,948	5,979	**594**	620	2,100	1,082	746	7,832	2,877
dbl	3,918	508	1010	934	**14**	25	29	13,101	23,106
Enr	875	124	**26**	27	114	55	34	577	7,067
EuA	1,663	1,419	**34**	36	179	71	35	254	458
fac	84	16	4	4	20	4	**3**	329	17
Gnu	1,558	**445**	638	637	583	593	739	6,139	11,118

Table 4.3: Relative performance $r_p = \frac{steps}{opt}$ of expected cost under the unit cost model

	AL	FS	SS-S	SS-R	SJ_0	$SJ_{0.5}$	SJ_1	Avra-W	Avra-S
GrQ	68.8	18.8	8.0	7.6	1.4	**1.0**	1.1	91.8	715.6
ama	11.7	10.1	**1.0**	1.0	3.5	1.8	1.3	13.2	4.8
dbl	279.9	36.3	72.1	66.7	**1.0**	1.8	2.1	935.8	1,650.4
Enr	33.7	4.8	**1.0**	1.0	4.4	2.1	1.3	22.2	271.8
EuA	48.9	41.7	**1.0**	1.1	5.3	2.1	1.0	7.5	13.5
fac	28.0	5.3	1.3	1.3	6.7	1.3	**1.0**	109.7	5.7
Gnu	3.5	**1.0**	1.4	1.4	1.3	1.3	1.7	13.8	25.0

4.5.6 Algorithm performance on real-world graphs with linear cost model

Tab. 4.4 and Tab. 4.5 give the expected cost of finding a degree λ node under the linear cost model. In the assortative graphs GrQc and dblp SAWJ again outperforms the competing algorithms under optimal β parameterization. However in the facebook graph Avra-S is dominant suggesting under the linear cost model Avra-S is extremely efficient in dense powerlaw graphs. In fact on the facebook graph SS-S, SS-R, and Avra-W all outperform SAWJ under the linear cost model which may be due

to both $\alpha \approx 0$ and the linear cost model. On the disassortative graphs amazon, Enron, EuAll, and Gnutella the absolute relative error of the optimal β parameterizations of SAWJ under the linear cost model is $r_p \leq 1.3$, performing comparatively or better then the two star sampling algorithms SS-R and SS-S. The algorithms AL, FS, and Avra-W all perform significantly worse then their competitors under the linear cost model.

Table 4.4: Expected cost in the linear cost model

	AL	FS	SS-S	SS-R	SJ_0	$SJ_{0.5}$	SJ_1	Avra-W	Avra-S
GrQ	2,820	22,163	393	397	284	190	**185**	734	5,725
ama	38,624	66,622	3,822	3,922	14,968	6,235	3,645	7,823	**2,877**
dbl	25,354	10,325	7,024	6,702	**2,135**	2,602	2,495	13,101	23,106
Enr	9,291	15,406	**274**	295	6,834	2,121	292	577	7,067
EuA	4,708	653,819	125	132	48,260	18,744	**95**	254	458
fac	3,326	1,622	119	126	4,187	769	725	329	**17**
Gnu	7,502	5,793	3,425	3,829	4,460	4,046	**3,072**	6,139	11,117

Table 4.5: Relative performance $r_p = \frac{steps}{opt}$ of expected cost under the linear cost model

	AS	FS	SS-S	SS-R	SJ_0	$SJ_{0.5}$	SJ_1	Avra-W	Avra-S
GrQ	15.2	119.8	2.1	2.1	1.5	1.0	**1.0**	4.0	30.9
ama	13.4	23.2	1.3	1.4	5.2	2.2	1.3	2.7	**1.0**
dbl	11.9	4.8	3.3	3.1	**1.0**	1.2	1.2	6.1	10.8
Enr	33.9	56.2	**1.0**	1.1	24.9	7.7	1.1	2.1	25.8
EuA	50.0	6,882.3	1.3	1.4	508.0	197.3	**1.0**	2.7	4.8
fac	195.6	95.4	7.0	7.4	246.3	45.2	42.6	19.4	**1.0**
Gnu	3.4	1.9	1.1	1.2	1.5	1.3	**1.0**	2.0	3.6

4.6 Conclusions

This chapter presented the SAWJ algorithm, designed to find maximum degree nodes in large assortative graphs. The algorithm was shown to perform well relative to SS-R, SS-S, AL, FS, Avra-S, and Avra-W both for a wide range of AER graphs and for six real-world large graphs from the Stanford SNAP dataset, although under the linear model Avra-S appears to perform significantly better on dense Powerlaw graphs. Additionally this chapter gave Markov models allowing the approximate cost for a simplified version of SAWJ, WJ, to find a max degree node for both an arbitrary graph and an AER graph to be computed, and the approximations are shown to be reasonably accurate.

4.7 Appendix

Proof of Prop. 8

Proof. The inverse of $\mathbf{C}_n$ and its square root can be shown to be

$$\mathbf{C}_n^{-1} = \frac{1}{(n-2)(s-1)s(\alpha^2-1)} \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix}, \quad \mathbf{C}_n^{-1/2} = \frac{1}{2} \begin{bmatrix} b+c & -b+c \\ -b+c & b+c \end{bmatrix} \tag{4.64}$$

where $b = \frac{1}{\sqrt{(n-2)(s-1)s(\alpha-1)}}$ and $c = \frac{1}{\sqrt{(2-n)(s-1)s(\alpha+1)}}$.

Therefore the standardization of $\mathbf{W}_n$ is

$$\begin{aligned} \tilde{\mathbf{W}}_n &= \mathbf{C}_n^{-1/2}(\mathbf{W}_n - \boldsymbol{\kappa}_n) && (4.65)\\ &= \frac{1}{2} \begin{bmatrix} b+c & -b+c \\ -b+c & b+c \end{bmatrix} \begin{bmatrix} \mathsf{A}_n - (1+(n-2)s) \\ \mathsf{B}_n - (1+(n-2)s) \end{bmatrix} && (4.66)\\ &= \frac{1}{2} \begin{bmatrix} (b+c)(\mathsf{A}_n - (1+(n-2)s)) + (c-b)(\mathsf{B}_n - (1+(n-2)s)) \\ (c-b)(\mathsf{A}_n - (1+(n-2)s)) + (b+c)(\mathsf{B}_n - (1+(n-2)s)) \end{bmatrix} && (4.67) \end{aligned}$$

and the standardization function Eq. (4.17) for each element of $\tilde{\mathbf{W}}_n$ is

$$\tilde{\mathsf{A}}_n = h_{\tilde{\mathsf{A}}_n}(\mathsf{A}_n, \mathsf{B}_n) = \frac{1}{2}(b+c)(\mathsf{A}_n - (1+(n-2)s)) + \frac{1}{2}(c-b)(\mathsf{B}_n - (1+(n-2)s)) \tag{4.68}$$

$$\tilde{\mathsf{B}}_n = h_{\tilde{\mathsf{B}}_n}(\mathsf{A}_n, \mathsf{B}_n) = \frac{1}{2}(c-b)(\mathsf{A}_n - (1+(n-2)s)) + \frac{1}{2}(b+c)(\mathsf{B}_n - (1+(n-2)s)) \tag{4.69}$$

Letting $\tilde{\mathbf{W}}_n = [\tilde{\mathsf{A}}_n \ \tilde{\mathsf{B}}_n]^\mathsf{T}$ it follows that the mean of each element is zero since by Lem. 8 $\mathbb{E}[\mathsf{A}_n] = \mathbb{E}[\mathsf{B}_n] = 1 + (n-2)s$ and,

$$\begin{aligned} \mathbb{E}[\tilde{\mathsf{A}}_n] &= \mathbb{E}[\tilde{\mathsf{B}}_n] && (4.70)\\ &= \mathbb{E}\left[\frac{1}{2}(c-b)(\mathsf{A}_n - (1+(n-2)s)) + \frac{1}{2}(b+c)(\mathsf{B}_n - (1+(n-2)s))\right] && (4.71)\\ &= \frac{1}{2}(c-b)(\mathbb{E}[\mathsf{A}_n] - (1+(n-2)s)) + \frac{1}{2}(b+c)(\mathbb{E}[\mathsf{B}_n] - (1+(n-2)s)) && (4.72)\\ &= 0 && (4.73) \end{aligned}$$

which gives $\mathbb{E}[\tilde{\mathbf{W}}_n] = [0,0]^T$.

The covariance, $Cov(\tilde{\mathsf{A}}_n, \tilde{\mathsf{A}}_n)$ and $Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{B}}_n)$ are

$$\begin{aligned} Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{B}}_n) &= Cov(\tilde{\mathsf{A}}_n, \tilde{\mathsf{A}}_n) && (4.74)\\ &= Cov((b+c)\mathsf{A}_n + (c-b)\mathsf{B}_n, (b+c)\mathsf{A}_n + (c-b)\mathsf{B}_n) && (4.75)\\ &= (b+c)^2 Cov(\mathsf{A}_n, \mathsf{A}_n) + 2(c^2-b^2)Cov(\mathsf{A}_n, \mathsf{B}_n) + M^2 Cov(\mathsf{B}_n, \mathsf{B}_n) && (4.76)\\ &= (b+c)^2 Var(\mathsf{A}_n) + 2(c^2-b^2)Cov(\mathsf{A}_n, \mathsf{B}_n) + (c-b)^2 Var(\mathsf{B}_n) && (4.77) \end{aligned}$$

Substituting $(b+c)^2$, $(c-b)^2$, and (c^2-b^2) into $Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{B}}_n)$ and recalling from Lem. 8 that $Var(\mathsf{A}_n) = Var(\mathsf{B}_n)(n-2)s(1-s)$ and $Cov(\mathsf{A}_n, \mathsf{B}_n) = (n-2)s(1-s)\alpha$ simplifying $Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{B}}_n)$ gives,

$$\begin{aligned} Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{B}}_n) &= \frac{1}{2}\left[\frac{1}{|(1-\alpha)|} + \frac{1}{|(1+\alpha)|}\right] + \frac{1}{2}\left[\frac{-\alpha}{|(1-\alpha)|} + \frac{\alpha}{|(1+\alpha)|}\right] && (4.78)\\ &= \frac{1}{2}\left[\frac{1-\alpha}{|(1-\alpha)|} + \frac{1+\alpha}{|(1+\alpha)|}\right] && (4.79)\\ &= 1 && (4.80) \end{aligned}$$

Similarly the covariance between $\tilde{\mathsf{A}}_n$ and $\tilde{\mathsf{B}}_n$ is

$$\begin{aligned} Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{A}}_n) &= Cov(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) && (4.81)\\ &= 2(c^2-b^2)Var(\mathsf{A}_n) + (b+c)^2 Cov(\mathsf{A}_n, \mathsf{B}_n) + (c-b)^2 Cov(\mathsf{B}_n, \mathsf{A}_n) && (4.82) \end{aligned}$$

and since $-1 < \alpha < 1$, substituting c^2-b^2, $(b+c)^2$, and $(c-b)^2$ into $Cov(\tilde{\mathsf{B}}_n, \tilde{\mathsf{A}}_n)$ gives

$$\begin{aligned} Cov(\tilde{\mathsf{A}}_n, \tilde{\mathsf{B}}_n) &= \frac{1}{2}\left(\frac{-1}{|\alpha-1|} + \frac{1}{|\alpha+1|}\right) + \frac{1}{2}\left(\frac{\alpha}{|\alpha-1|} + \frac{\alpha}{|\alpha+1|}\right) && (4.83)\\ &= 0 && (4.84) \end{aligned}$$

which gives $Cov(\tilde{\mathbf{W}}_n) = \mathbf{I}_2$. □

Chapter 5: On the number of star samples to find a vertex or edge with given degree in a graph

5.1 Introduction

5.1.1 Motivation

This chapter considers the tasks of finding a vertex of a specified degree, say k, or finding an edge with specified degrees, say $\{j, k\}$, in a large undirected graph. For example, if one knows the maximum or average degree of a graph then one may seek to find a vertex with that degree. This analysis assumes the graph may be queried by specifying a vertex index, and the graph will return, at unit computational and communication cost, the *star sample* (SS) for that vertex. The SS consists of the vertex, its (one-hop) neighborhood, and the degrees (but not the neighbors) of each vertex in the neighborhood. Such samples are natural in many real-world contexts, e.g., social networks, where querying a person (vertex) reveals their "friends", and (in some cases) the number of friends held by each friend. This capability for querying the graph is employed by collecting a random star samples (SS-R), wherein a random vertex is repeatedly identified, with replacement, independently and uniformly at random, and the SS associated with this vertex is reviewed. This process is repeated until a SS containing a vertex (either as the center or a neighbor) or an edge with the specified degrees is found. The focus of this chapter is on identifying the probability that a SS-R contains the vertex or edge of interest, with the inverse of this probability the expected (Geometrically distributed) number of samples required. This focus is further restricted to the case of classic Erdős-Rényi (ER) random graph family with order n, wherein each of the possible $\binom{n}{2}$ potential edges is added independently with probability s. To help in this analysis this chapter introduces a modified construction of the classic construction so that edges are *placed* independently and at random, yielding a multigraph.

Since Chap. 6 introduces a simpler and more accurate analysis of the probability of SS-R sampling a vertex of interest, this chapters main contributions are the introduction of the modified ER graph and the resulting analysis which justifies the approximation in Def. 16 of the probability that a degree j node contains a degree k node in its neighborhood. Specifically Thm. 10 states that the fraction of multiple edges in a modified ER graph is asymptotically small, Thm. 11 states that the degree distribution of the classic and modified ER graph are asymptotically close up to the second moment for small s. Finally Thm. 12 shows that the approximation given in Def. 16 holds conditionally in

modified ER graphs Thm. 12, and it follows that if Thm. 10 and Thm. 11 hold the approximation also holds asymptotically under the same conditions for classic ER graphs.

One may view SS-R as an intermediate between the two "graph search extremes" of *random sampling* (RS) and *random walks* (RW). In RS a vertex is repeatedly drawn uniformly at random, either with or without replacement, until the vertex with desired degree is found. Under a RW the next vertex is repeatedly chosen to be a neighbor of the current vertex, according to a neighbor selection rule. These two approaches are opposites in the sense that RS makes no use of graph structure, while the RW relies entirely on this structure. Unfortunately, the RW suffers the drawback of locality, restricting itself to a path through the graph. SS-R is intermediate in the sense that it leverages the local information at a given vertex (like the RW), but without RW's locality deficiency. As such, the estimated number of samples of a SS-R is of potential value. Recall that Chaps. 2 to 4 have in part studied the performance of a biased RW (BRW) to find a maximum degree vertex and have found that for positively (negatively) assortative graphs BRWs are superior (inferior) to SS-R in finding maximum degree vertices.

5.1.2 Related work

Significant work has been done on star and survey sampling in graphs. SS-R is discussed in Kolaczyk [2009] within the more general concept of snowball sampling, a classic survey sampling technique Frank [1977], Goodman [1961]. Given a population of size n, define $y = y_1 + \cdots + y_n$ as the total value of this population and $\mathbf{w} = [w_1, \ldots, w_n]$ as the probability of sampling each member of the population. If one takes a size z sample of this population, the Horvitz-Thompson (HT) estimator Frank [1977] gives an approximation of y. In the context of a graph, the HT estimator gives the probability a given vertex u is included in a size z star sample, and may be derived via the principle of inclusion and exclusion Kolaczyk [2009]. In contrast to Kolaczyk [2009], the focus in this chapter is on estimating the number of SS-R required to find *any* vertex or edge with specified degree(s).

5.1.3 Contributions and outline

In this chapter Sec. 5.2 defines the true probability and its combinatorial approximation that an SS-R contains a vertex or edge of interest. The approximation is a function of pertinent summary statistics of the graph, namely the number of degree k vertices and the number of edges with degrees $\{j, k\}$, and is therefore more easily computed than the true probability. In Sec. 5.3 assess the approximation's accuracy for the ER graph by first defining a modification of the ER construction, wherein each edge is *placed* uniformly at random, yielding a multigraph. This chapter gives three results. Thm. 10

establishes that the fraction of multiple edges is exponentially small in the ER edge probability s, i.e., the multigraph is, with high probability, "graph-like". Thm. 11 establishes the expected degree of a random vertex under the classic and modified ER constructions are equal, and the variances of the random degree have a ratio close to one for small s, i.e., the multigraph is, with high probability, "ER-like". Thm. 12 shows the ratio of the true probability over the approximation is near one, i.e., the approximation is accurate for modified ER multigraphs $\tilde{G}_\epsilon$. Finally this chapter argues that as well as G_ϵ the approximation is accurate for classic ER graphs. Sec. 5.4 gives numerical results for synthetic ER graphs and "real-world" SNAP graphs; the approximation is excellent for ER graphs and mixed for SNAP graphs. Sec. 5.5 holds a brief conclusion, and the proof of Thm. 10 is in Sec. 5.6.

5.2 Star sampling with replacement

5.2.1 Notation.

General notation. Write $[n]^+ \equiv \{1, \ldots, n\}$, for $n \in \mathbb{N}$ and let $M \equiv \binom{n}{2}$. For any $a \in [0,1]$, let $\bar{a} \equiv 1 - a$ be its "complement". Random variables (RVs) are denoted in sans-serif font, e.g., u, X, with probability $\mathbb{P}(\cdot)$, expectation $\mathbb{E}[\cdot]$, variance $\mathrm{Var}(\cdot)$, and probability generating function $\phi(\cdot)$. Let $\mathsf{u} \sim Uni(\mathcal{V})$ denote a Uniform RV drawn from set $\mathcal{V}$, and $\mathsf{m} \sim \mathrm{Bin(M, s)}$ denote a binomial RV with parameters (M, s).

Graph notation. Consider an undirected labeled simple graph $G = (\mathcal{V}, \mathcal{E})$ with order $n \equiv |\mathcal{V}|$ (order) and size $m \equiv |\mathcal{E}|$.

- Let d_v denote the degree of vertex v, $\mathcal{D} \equiv \bigcup_{v \in \mathcal{V}} d_v$ the set of degrees in the graph, and $\lambda \equiv \max(\mathcal{D})$ the max degree.
- Let $\mathcal{V}_k \equiv \{v \in \mathcal{V} | d_v = k\}$ be the set of degree k vertices, and $n_k \equiv |\mathcal{V}_k|$ the number of such vertices. Note $\sum_{k \in \mathcal{D}} n_k = n$, and $\mu \equiv \sum_{k \in \mathcal{D}} k \frac{n_k}{n}$ is the average degree.
- Let $d(e)$ be the degrees of the endpoints of edge $e \in \mathcal{E}$, e.g., $d(e) = \{j, k\}$ means edge e has vertices of degrees $\{j, k\}$. Let $\mathcal{E}_k \equiv \{e \in \mathcal{E} | k \in d(e)\}$ be the set of edges with a vertex of degree k, with $m_k \equiv |\mathcal{E}_k|$ as the number of such edges. An edge in $\mathcal{E}_k$ will be called a degree k edge. If $d(e) = \{j, k\}$ for $k \neq j$ then e is both a degree k and a degree j edge, for $k = j$ then e counts as 2 degree k edges.
- Let $\mathcal{E}_{j,k} \equiv \{e \in \mathcal{E} | d(e) = \{j, k\}\}$ be the edges with degrees $\{j, k\}$, and let $m_{j,k} = |\mathcal{E}_{j,k}|$ be the number of such edges; $\mathcal{E}_{k,k}$ ($m_{k,k}$) is the set (number) of edges with both endpoints of degree k.

Edges in set $\mathcal{E}_{j,k}$ are called degree $\{j,k\}$ edges.

- Each edge is an unordered pair of distinct vertices, denoted $e = uv = vu$, for $u, v \in \mathcal{V}^2$. The *edge* neighborhood of vertex $v \in \mathcal{V}$, i.e., edges in $\mathcal{E}$ adjacent to v, is denoted $\mathcal{N}_v$.

5.2.2 Overview of approach

Star sampling a graph with replacement, to find a particular vertex or edge, refers to the following procedure: repeatedly select a vertex $\mathsf{u} \sim Uni(\mathcal{V})$ uniformly at random, check if the edge neighborhood $\mathcal{N}_{\mathsf{u}}$ of the sample contains a vertex or edge of interest, and stop when found. This chapter considers two specific objectives in this context: *i*) find a degree k edge (at least one of the two degrees of the edge is k), and *ii*) find an edge with degrees $\{j,k\}$. In star sampling finding a degree k vertex (i.e., in $\mathcal{V}_k$) is equivalent to finding a degree k edge.

Since this chapter considers SS-R i.e. sampling with replacement, the number of samples is a Geometric RV with some parameter p and therefore expected value $1/p$. When the objective is to find a degree k edge (degree $\{j,k\}$ edge), the parameter is $\bar{u}_k$ ($\tilde{U}_{j,k}$), where $\bar{u}_k$ ($\tilde{U}_{j,k}$) is the probability a star sample contains a degree k (degree $\{j,k\}$) edge in its neighborhood.

The approach taken in this chapter is to derive an approximation to these parameters (e.g., y_k below) and to define a bound on the accuracy of the approximation (e.g., ϵ_k below). The approximation is a simple combinatorial quantity that can be computed from knowledge of the degree distribution ($n_j/n, j \in \mathcal{D}$), the number of edges of interest (e.g., $m_{j,k}$ below), and the number of degree j edges in the graph m_j. The bound, however, can only be computed with knowledge of the entire graph. This chapter takes two steps in an attempt to address this: *i*) The accuracy of the approximation numerically for both synthetic and real-world graphs is computed, and *ii*) the accuracy of the approximation analytically for the special case of ER graphs is studied.

5.2.3 Star sampling with replacement to find a degree k edge

Consider taking the m_j edges $\mathcal{E}_j$ from the graph G and drawing a random sample without replacement from $\mathcal{E}_j$ of size j. There are $\binom{m_j}{j}$ distinct samples of these edges, and $\binom{m_j - m_{j,k}}{j}$ distinct samples that do not contain a degree k edge.

Definition 16. *Define the* $|\mathcal{D}| \times |\mathcal{D}|$ *matrices* $\mathbf{U} = (U_{j,k})$ *and* $\mathbf{Y} = (Y_{j,k})$ *where* $\mathbf{Y}_{j,j} \equiv 0$*, and, for distinct* $(j,k) \in \mathcal{D}^2$

$$U_{j,k} \equiv \mathbb{P}(\mathcal{N}_{\mathsf{u}_j} \cap \mathcal{E}_k = \emptyset),\ Y_{j,k} \equiv \binom{m_j - m_{j,k}}{k} \Big/ \binom{m_j}{j}. \tag{5.1}$$

$U_{j,k}$ is the probability a randomly sampled degree j vertex, u_j, doesn't contain a degree k edge in its neighborhood, and $Y_{j,k}$ is the probability a random j-sample (without replacement) of m_j edges doesn't contain a degree k edge. Let $\mathbf{E}$ be the $|\mathcal{D}| \times |\mathcal{D}|$ deviation matrix with entries $E_{j,k} \equiv |U_{j,k} - Y_{j,k}|$.

As $\mathsf{u}_j \sim Uni(\mathcal{V}_j)$ is a randomly selected degree j vertex,

$$U_{j,k} = \frac{1}{n_j} \sum_{v \in \mathcal{V}_j} \mathbf{1}(\mathcal{N}_v \cap \mathcal{E}_k = \emptyset) \tag{5.2}$$

is the fraction of degree j vertices not adjacent to a degree k edge. Observe $U_{j,j} = 0$ for each $j \in \mathcal{D}$, by construction.

Definition 17. *Define the $|\mathcal{D}|$-vectors $\mathbf{u} = (u_k, k \in \mathcal{D})$, $\mathbf{y} = (y_k, k \in \mathcal{D})$, and $\boldsymbol{\epsilon} = (\epsilon_k, k \in \mathcal{D})$ where*

$$u_k \equiv \sum_{j \in \mathcal{D}} U_{j,k} \frac{n_j}{n},\ y_k \equiv \sum_{j \in \mathcal{D}} Y_{j,k} \frac{n_j}{n},\ \epsilon_k \equiv \sum_{j \in \mathcal{D}} E_{j,k} \frac{n_j}{n}, \tag{5.3}$$

are weighted combinations of the k-column of the matrices $\mathbf{U}, \mathbf{Y}, \mathbf{E}$, with weights corresponding to the degree distribution.

The probability the edge neighborhood of a randomly selected vertex, $\mathsf{u} \sim Uni(\mathcal{V})$, doesn't contain a degree k edge is

$$u_k \equiv \mathbb{P}(\mathcal{N}_{\mathsf{u}} \cap \mathcal{E}_k = \emptyset) = \frac{1}{n} \sum_{v \in \mathcal{V}} \mathbf{1}(\mathcal{N}_v \cap \mathcal{E}_k = \emptyset). \tag{5.4}$$

Proposition 11. *For any $k \in \mathcal{D}$, the probability a star sample contains a degree k vertex, $\bar{u}_k$, is within ϵ_k of $\bar{y}_k$:*

$$|\bar{u}_k - \bar{y}_k| = |u_k - y_k| \leq \epsilon_k. \tag{5.5}$$

Consequently, the expected number of star samples to find such a vertex, $1/\bar{u}_k$, is bounded by

$$\frac{1}{\bar{y}_k + \epsilon_k} \leq \frac{1}{\bar{u}_k} \leq \frac{1}{\bar{y}_k - \epsilon_k}. \tag{5.6}$$

Proof. $|\bar{u}_k - \bar{y}_k| = |u_k - y_k|$

$$\begin{aligned} &= \left| \sum_{j \in \mathcal{D}} U_{j,k} \frac{n_j}{n} - \sum_{j \in \mathcal{D}} Y_{j,k} \frac{n_j}{n} \right| = \left| \sum_{j \in \mathcal{D}} (U_{j,k} - Y_{j,k}) \frac{n_j}{n} \right| \\ &\leq \sum_{j \in \mathcal{D}} |U_{j,k} - Y_{j,k}| \frac{n_j}{n} \leq \sum_{j \in \mathcal{D}} E_{j,k} \frac{n_j}{n} = \epsilon_k \end{aligned} \tag{5.7}$$

It follows that $\bar{u}_k \in (\bar{y}_k - \epsilon_k, \bar{y}_k + \epsilon_k)$, and as such Eq. (5.6) follows from the fact that the number

of star samples is a Geometric RV with parameter $\bar{u}_k$, and expected value $1/\bar{u}_k$. □

The combinatorial quantities $Y_{j,k}$ from Def. 16 and y_k from Def. 17 may be bounded and approximated as follows.

Proposition 12. *The quantity $Y_{j,k}$ obeys $Y^{\mathrm{L}}_{j,k} \leq Y_{j,k} \leq Y^{\mathrm{U}}_{j,k}$, for $j + m_{j,k} < m_j$, where*

$$Y^{\mathrm{L}}_{j,k} = \left(1 - \frac{m_{j,k} + (j-1)}{m_j - (j-1)}\right)^j, \; Y^{\mathrm{U}}_{j,k} = \left(1 - \frac{m_{j,k}}{m_j}\right)^j, \tag{5.8}$$

yielding bounds $y^{\mathrm{L}}_k \leq y_k \leq y^{\mathrm{U}}_k$ where

$$y^{\mathrm{L}}_k = \sum_{j \in \mathcal{D}} Y^{\mathrm{L}}_{j,k} \frac{n_j}{n}, \quad y^{\mathrm{U}}_k = \sum_{j \in \mathcal{D}} Y^{\mathrm{U}}_{j,k} \frac{n_j}{n}. \tag{5.9}$$

Finally, if $m_{j,k} \ll m_k$ then:

$$y^{\mathrm{U}}_k \approx 1 - \frac{m_{j,k}}{m_j}\mu. \tag{5.10}$$

The approximation in Eq. (5.10) is not a bound on y_k, as Bernoulli's inequality gives a *lower* bound on $Y^{\mathrm{U}}_{j,k}$.

Proof. First observe

$$Y_{j,k} = \frac{\binom{m_j - m_{j,k}}{j}}{\binom{m_j}{j}} = \prod_{l=0}^{j-1} \left(1 - \frac{m_{j,k} + l}{m_j - l}\right). \tag{5.11}$$

The terms in the product are decreasing in l and as such the product is bounded below by $Y^{\mathrm{L}}_{j,k}$ and above by $Y^{\mathrm{U}}_{j,k}$. To see Eq. (5.10), recall Bernoulli's inequality $(1-x)^t \geq 1 - xt$, with the inequality an approximation valid for small x. In the current context, if $m_{j,k} \ll m_j$ then $Y^{\mathrm{U}}_{j,k} \approx 1 - j\frac{m_{j,k}}{m_j}$, and thus

$$y^{\mathrm{U}}_k \approx \sum_{j \in \mathcal{D}} \left(1 - j\frac{m_{j,k}}{m_j}\right) \frac{n_j}{n} = 1 - \frac{m_{j,k}}{m_j} \sum_{j \in \mathcal{D}} j \frac{n_j}{n}. \tag{5.12}$$

□

5.2.4 Star sampling with replacement to find a $\{j,k\}$ edge

The definitions in Sec. 5.2.3 can be extended in a natural way for the objective of finding a $\{j,k\}$ edge.

Definition 18. *Define $\tilde{U}_{j,k} = \tilde{U}_{j,k} \equiv \mathbb{P}(\mathcal{N}(\mathsf{u}) \cap \mathcal{E}_{j,k} \neq \emptyset)$, for $(j,k) \in \mathcal{D}^2$, as the probability the neighborhood of a randomly selected vertex contains a degree $\{j,k\}$ edge.*

Note $\tilde{U}_{j,k} = \tilde{U}_{k,j}$ is symmetric, but $U_{j,k}$ need not equal $U_{k,j}$.

Lemma 14. *The quantity $\tilde{U}_{j,k}$ in Def. 18 is expressible in terms of $U_{j,k}$ and $U_{k,j}$ from Def. 16:* $\tilde{U}_{j,k} = \bar{U}_{j,k}\frac{n_j}{n} + \bar{U}_{k,j}\frac{n_k}{n}$.

Proof. Condition on the degree $l \in \mathcal{D}$ of the selected vertex, and observe the event that the star sample of a degree l vertex contains a $\{j,k\}$ edge requires either $l = j$ or $l = k$.

$$\begin{aligned}
\tilde{U}_{j,k} &\equiv \mathbb{P}(\mathcal{N}(\mathsf{u}) \cap \mathcal{E}_{j,k} \neq \emptyset) \\
&= \sum_{l \in \mathcal{D}} \mathbb{P}(\mathcal{N}(\mathsf{u}_l) \cap \mathcal{E}_{j,k} \neq \emptyset)\frac{n_l}{n} \\
&= \mathbb{P}(\mathcal{N}(\mathsf{u}_j) \cap \mathcal{E}_k \neq \emptyset)\frac{n_j}{n} + \mathbb{P}(\mathcal{N}(\mathsf{u}_k) \cap \mathcal{E}_j = \emptyset)\frac{n_k}{n} \\
&= \bar{U}_{j,k}\frac{n_j}{n} + \bar{U}_{k,j}\frac{n_k}{n} \qquad (5.13)
\end{aligned}$$

□

The definition of $\tilde{Y}_{j,k}$ below is inspired by $\tilde{U}_{j,k}$ in Lem. 14.

Definition 19. *Define $\tilde{Y}_{j,k}$ as the probability a random sized random sample of edges contains a certain degree edge. In particular, the random sample size is j (w.p. n_j/n), k (w.p. n_k/n), or 0 (w.p. $1 - n_j/n - n_k/n$), where the edge of interest is a degree k (j) edge for a size j (k) sample, respectively:*

$$\tilde{Y}_{j,k} = \tilde{Y}_{k,j} \equiv \bar{Y}_{j,k}\frac{n_j}{n} + \bar{Y}_{k,j}\frac{n_k}{n}. \qquad (5.14)$$

Define the deviation $\tilde{E}_{j,k} = \tilde{E}_{k,j} \equiv |\tilde{U}_{j,k} - \tilde{Y}_{j,k}|$.

Proposition 13. *For any $(j,k) \in \mathcal{D}^2$, the deviation $\tilde{E}_{j,k}$ of $\tilde{U}_{j,k}$ from $\tilde{Y}_{j,k}$ is upper bounded by $\tilde{E}^{\mathrm{U}}_{j,k} = \tilde{E}^{\mathrm{U}}_{k,j}$:*

$$\tilde{E}_{j,k} \leq \tilde{E}^{\mathrm{U}}_{j,k} \equiv E_{j,k}\frac{n_j}{n} + E_{k,j}\frac{n_k}{n}. \qquad (5.15)$$

Consequently, the expected number of star samples to find such a edge, $1/\tilde{U}_{j,k}$, is bounded by

$$\frac{1}{\tilde{Y}_{j,k} + \tilde{E}^{\mathrm{U}}_{j,k}} \leq \frac{1}{\tilde{Y}_{j,k} + \tilde{E}_{j,k}} \leq \frac{1}{\tilde{U}_{j,k}} \leq \frac{1}{\tilde{Y}_{j,k} - \tilde{E}_{j,k}} \leq \frac{1}{\tilde{Y}_{j,k} - \tilde{E}^{\mathrm{U}}_{j,k}}. \qquad (5.16)$$

Proof. Simply substitute the appropriate definitions:

$$\begin{aligned}
\tilde{E}_{j,k} &= |\tilde{U}_{j,k} - \tilde{Y}_{j,k}| \\
&= \left|\left(\bar{U}_{j,k}\frac{n_j}{n} + \bar{U}_{k,j}\frac{n_k}{n}\right) - \left(\bar{Y}_{j,k}\frac{n_j}{n} + \bar{Y}_{k,j}\frac{n_k}{n}\right)\right| \\
&= \left|\left(\bar{U}_{j,k} - \bar{Y}_{j,k}\right)\frac{n_j}{n} + \left(\bar{U}_{k,j} - \bar{Y}_{k,j}\right)\frac{n_k}{n}\right| \\
&= \left|\left(Y_{j,k} - U_{j,k}\right)\frac{n_j}{n} + \left(Y_{k,j} - U_{k,j}\right)\frac{n_k}{n}\right| \\
&\leq |Y_{j,k} - U_{j,k}|\frac{n_j}{n} + |Y_{k,j} - U_{k,j}|\frac{n_k}{n} \\
&= E_{j,k}\frac{n_j}{n} + E_{k,j}\frac{n_k}{n} = \tilde{E}^{\mathrm{U}}_{j,k}
\end{aligned} \tag{5.17}$$

□

The next corollary applies the bounds in Prop. 12 to $\tilde{Y}_{j,k}$.

Corollary 3. *$\tilde{Y}_{j,k}$ obeys $\tilde{Y}^{\mathrm{L}}_{j,k} \leq \tilde{Y}_{j,k} \leq Y^{\mathrm{U}}_{j,k}$, where:*

$$\begin{aligned}
\tilde{Y}^{\mathrm{L}}_{j,k} &= \left(1 - Y^{\mathrm{U}}_{j,k}\right)\frac{n_j}{n} + \left(1 - Y^{\mathrm{U}}_{k,j}\right)\frac{n_k}{n} \\
\tilde{Y}^{\mathrm{U}}_{j,k} &= \left(1 - Y^{\mathrm{L}}_{j,k}\right)\frac{n_j}{n} + \left(1 - Y^{\mathrm{L}}_{k,j}\right)\frac{n_k}{n}
\end{aligned} \tag{5.18}$$

Proof. The result follows from Def. 19 and Prop. 12. □

To summarize: the probabilities $\bar{u}_k$ and $\tilde{U}_{j,k}$ of finding an edge of interest in a star sample are approximated by the combinatorial quantities $\bar{y}_k$ and $\tilde{Y}_{j,k}$, which are in turn bounded by $(y^{\mathrm{L}}_k, y^{\mathrm{U}}_k)$ and $(\tilde{Y}^{\mathrm{L}}_{j,k}, \tilde{Y}^{\mathrm{U}}_{j,k})$, with error bounds $\epsilon_k = |\bar{u}_k - \bar{y}_k|$ and $\tilde{E}_{j,k} = |\tilde{U}_{j,k} - \tilde{Y}_{j,k}|$, respectively. Numerical evaluations of these quantities are given in Sec. 5.4, but the next section gives an analysis of the approximation on classic and modified ER graphs.

5.3 Star sampling with replacement in Erdős Rényi (ER) random graphs (RG)

The classic Erdős Rényi (ER) random graph (RG) construction with parameters (n, s) yields a RG where the $M \equiv \binom{n}{2}$ potential edges are added independently with probability s. The graph size is the RV $\mathsf{m} \sim \mathrm{Bin}(M, s)$. This construction is equivalent to Alg. 4, where the random size m is selected first, and the positions of the edges are selected at random.

Contrast the classic ER construction of Alg. 4 with the modified ER construction of Alg. 5, where the only change is that now the edge set is selected *with* replacement, instead of without. It follows that there may be multiple edges in $\tilde{G}_\epsilon$. The motivation for this construction is that the random

Algorithm 4 "Classic" Erdős Rényi (ER) random graph (RG)

1: **Input parameters** (n, s), $n \in \mathbb{N}$, $s \in (0, 1)$
2: Generate the RV $\mathsf{m} \sim \text{Bin}(M, s)$, the num. of edges
3: Select *without* replacement a set of m pairs of distinct vertices, and let the edge set $\mathcal{E}$ be the set of vertex pairs
4: **Return** the (simple) RG, $G_\epsilon = ([n], \mathcal{E})$

Algorithm 5 "Modified" Erdős Rényi (ER) random graph

1: **Input parameters** (n, s), $n \in \mathbb{N}$, $s \in (0, 1)$
2: Generate the RV $\tilde{\mathsf{m}} \sim \text{Bin}(M, s)$, the num. of edges
3: Select *with* replacement a set of $\tilde{\mathsf{m}}$ pairs of distinct vertices, and let the edge set $\tilde{\mathcal{E}}$ be the set of vertex pairs
4: **Return** the RG, $\tilde{G}_\epsilon = ([n], \tilde{\mathcal{E}})$

edges $\tilde{\mathcal{E}}$ in the modified ER construction are independent, while the random edges $\mathcal{E}$ in the classic ER construction are not.

5.3.1 Three properties of the modified ER RG $\tilde{G}_\epsilon$

The first property, Thm. 10, establishes the random fraction of multiple edges in $\tilde{G}_\epsilon$ is a RV that converges in probability to $1 - \mathrm{e}^{-s}$ as the order n grows to infinity. This means that for large n and small s (the case of typical interest), the fraction of multiple edges is small, and thus $\tilde{G}_\epsilon$ is *approximately* simple.

Theorem 10. *Let the RV $\tilde{\mathsf{y}}^{(n)}$ be the fraction of non-multiple edges in the modified ER RG, $\tilde{G}_\epsilon = (n, \tilde{\mathcal{E}})$. Then $\tilde{\mathsf{y}}^{(n)}$ converges in probability to e^{-s}, denoted $\tilde{\mathsf{y}}^{(n)} \xrightarrow{\text{P}} \mathrm{e}^{-s}$, as $n \uparrow \infty$.*

The proof of Thm. 10 is in Sec. 5.6. The second property, Thm. 11, establishes the RGs $(G_\epsilon, \tilde{G}_\epsilon)$ yield random degrees (for randomly selected vertices) with *i*) identical expectations (for all (n, s)), and *ii*) similar variances (for large n and small s).

Let $(\mathsf{z}, \tilde{\mathsf{z}})$ be the degrees of randomly selected vertex of the RGs $(G_\epsilon, \tilde{G}_\epsilon)$ constructed using Alg. 4 and Alg. 5, respectively and $\mathsf{z} \sim \text{Bin}(n-1, s)$, with $\mathbb{E}[\mathsf{z}] = (n-1)s$ and $\text{Var}(\mathsf{z}) = (n-1)s\bar{s}$. Let the RV $\tilde{\mathsf{m}} \sim \text{Bin}(M, s)$ be the number of unordered edges of $\tilde{G}_\epsilon$ constructed using Alg. 5, and the conditional RV $\tilde{\mathsf{z}}|\tilde{\mathsf{m}}$ be the conditional random degree of an arbitrary vertex of the random graph, conditioned on $\tilde{\mathsf{m}}$. Finally let the RV $\tilde{\mathsf{z}}$ be the unconditioned random degree of the vertex in $\tilde{G}_\epsilon$.

Theorem 11. *Conditioned on $\tilde{\mathsf{m}}$, the RV $\tilde{\mathsf{z}}|\tilde{\mathsf{m}}$ has a binomial distribution with $\tilde{\mathsf{m}}$ trials and trial success probability $2/n$: $\tilde{\mathsf{z}}|\tilde{\mathsf{m}} \sim \text{Bin}\left(\tilde{\mathsf{m}}, \frac{2}{n}\right)$. The unconditioned RV $\tilde{\mathsf{z}}$ has mean $\mathbb{E}[\tilde{\mathsf{z}}] = (n-1)s$ and*

variance $\mathrm{Var}(\tilde{\mathsf{z}}) = (n-1)s\bar{s}\left[\left(1-\frac{2}{n}\right)\frac{1}{\bar{s}}+\frac{2}{n}\right]$. *The means and variances of* z *and* $\tilde{\mathsf{z}}$ *have ratios*

$$\frac{\mathbb{E}[\tilde{\mathsf{z}}]}{\mathbb{E}[\mathsf{z}]} = 1, \quad \frac{\mathrm{Var}(\tilde{\mathsf{z}})}{\mathrm{Var}(\mathsf{z})} = \left(1-\frac{2}{n}\right)\frac{1}{\bar{s}}+\frac{2}{n}. \tag{5.19}$$

Note $\mathrm{Var}(\tilde{\mathsf{z}})/\mathrm{Var}(\mathsf{z}) \approx 1$ for n "large" and s "small".

Proof. The conditional RV $\tilde{\mathsf{z}}|\tilde{\mathsf{m}}$ has $\mathbb{E}[\tilde{\mathsf{z}}|\tilde{\mathsf{m}}] = \frac{2}{n}\tilde{\mathsf{m}}$ and $\mathrm{Var}(\tilde{\mathsf{z}}|\tilde{\mathsf{m}}) = \frac{2}{n}\left(1-\frac{2}{n}\right)\tilde{\mathsf{m}}$. The unconditioned RV $\tilde{\mathsf{z}}$ has mean

$$\mathbb{E}[\tilde{\mathsf{z}}] = \mathbb{E}[\mathbb{E}[\tilde{\mathsf{z}}|\tilde{\mathsf{m}}]] = \frac{2}{n}\mathbb{E}[\tilde{\mathsf{m}}] = \frac{2}{n}Ms = (n-1)s. \tag{5.20}$$

The RV $\tilde{\mathsf{z}}$ has variance given by the law of total variance:

$$\begin{aligned}
\mathrm{Var}(\tilde{\mathsf{z}}) &= \mathbb{E}[\mathrm{Var}(\tilde{\mathsf{z}}|\tilde{\mathsf{m}})] + \mathrm{Var}(\mathbb{E}[\tilde{\mathsf{z}}|\tilde{\mathsf{m}}]) \\
&= \mathbb{E}\left[\frac{2}{n}\left(1-\frac{2}{n}\right)\tilde{\mathsf{m}}\right] + \mathrm{Var}\left(\frac{2}{n}\tilde{\mathsf{m}}\right) \\
&= \frac{2}{n}\left(1-\frac{2}{n}\right)\mathbb{E}[\tilde{\mathsf{m}}] + \left(\frac{2}{n}\right)^2\mathrm{Var}(\tilde{\mathsf{m}}) \\
&= \frac{2}{n}\left(1-\frac{2}{n}\right)Ms + \left(\frac{2}{n}\right)^2 Ms\bar{s} \\
&= (n-1)s\bar{s}\left[\left(1-\frac{2}{n}\right)\frac{1}{\bar{s}}+\frac{2}{n}\right].
\end{aligned} \tag{5.21}$$

□

The third property asserts the modified ER has $U_{j,k} \approx Y_{j,k}$.

Theorem 12. *Under the modified ER construction*

$$\frac{U_{j,k}}{Y_{j,k}} = \left(\frac{w}{x}\right)^j \left(\frac{1-y\left(\frac{1}{x}\right)}{1-y\left(\frac{1}{w}\right)}\right)^{m_j-j} \left(\frac{1-y}{1-y\left(\frac{z}{w}\right)}\right)^{m_{j,k}} \tag{5.22}$$

where $w = \left(1-\frac{n_j-1}{2(n-1)}\right)\uparrow 1$, $x = \left(1-\frac{n_j-1}{2(n-1-n_k)}\right)\uparrow 1$, $y = \left(\frac{1}{n_j}\right)\downarrow 0$, $z = \left(1-\frac{n_k}{n-1}\right)\uparrow 1$*; whenever* n_j, n_k *are small relative to* n *and* $n \uparrow \infty$*, which implies* $U_{j,k} \approx Y_{j,k}$.

The proof of Thm. 12 is in Sec. 5.6.

5.4 Numerical Results

5.4.1 Synthetic graph results

For the numerical results 10 classic and modified ER graphs are generated per parameter pair (n, s) using Alg. 4 and Alg. 5 respectively where $n = 500$ and $s \in \{0.05 \ldots 0.95\}$. $U_{j,k}$ and $Y_{j,k}$ as defined

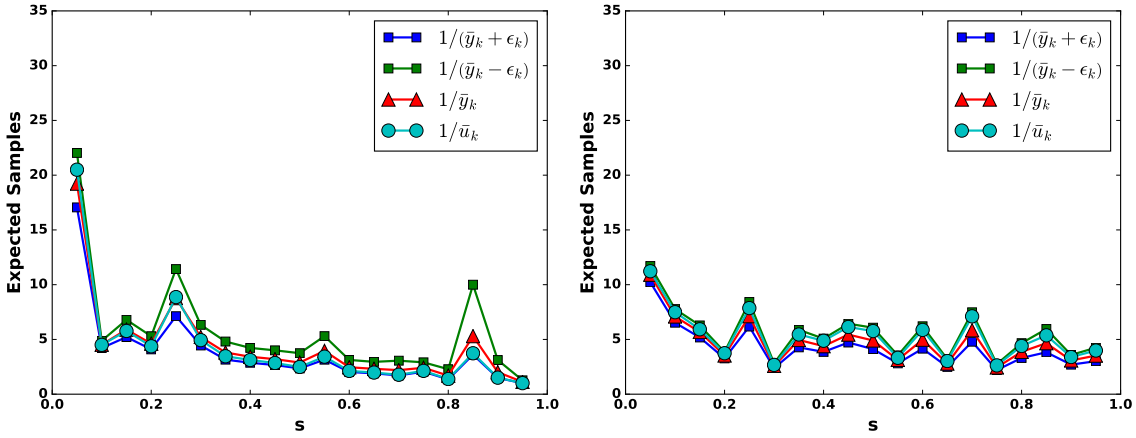

Figure 5.1: Expected number of SS-R to find a max. degree ($k = \lambda$) vertex (Prop. 11) for $n = 500$: classic ER graph G_ϵ (top), modified ER graph $\tilde{G}_\epsilon$ (bottom).

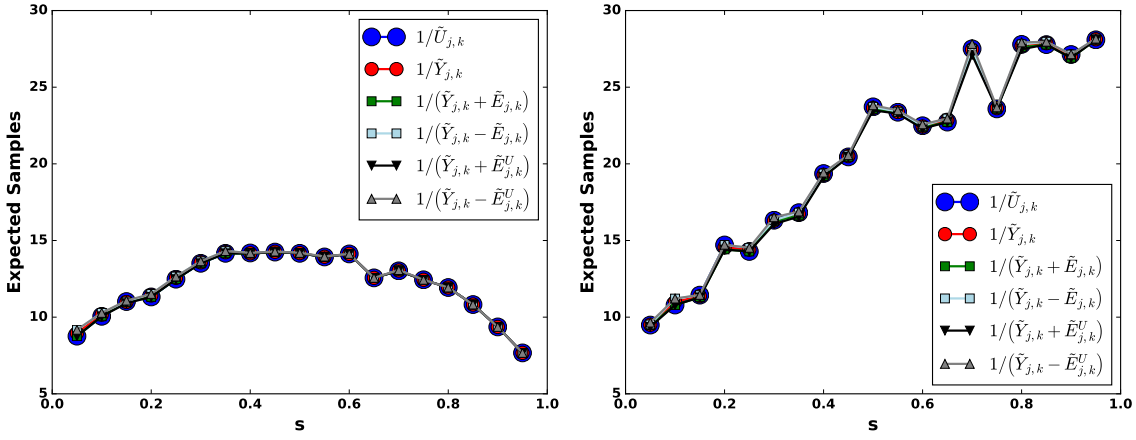

Figure 5.2: Expected number of SS-R to find a degree $\{j, k\}$ edge where $j = \lfloor ns \rfloor$, $k = \lfloor ns+4 \rfloor$, and $n = 500$ (Prop. 13): classic ER graph G_ϵ (top), modified ER graph $\tilde{G}_\epsilon$ (bottom).

in Def. 16 are averaged over the set of 10 graphs for each parameter pair (n, s).

Fig. 5.1 gives the numerical results for the expected number of star samples needed to find a degree k vertex (Prop. 12) in classic (top) and modified (bottom) ER graphs where $k = \lambda_\epsilon$ and $k = \tilde{\lambda}_\epsilon$ respectively. Fig. 5.1 suggests that the approximation of $1/\bar{u}_k$ by $1/\bar{y}_k$ is good in the classic ER case, slightly better in the modified ER case, and that the bounds given in Prop. 12 are fairly tight. Fig. 5.2 gives the numerical results for the expected number of star samples needed to find a (j, k) edge (Prop. 13) in the classic (Top) and modified (Bottom) ER graphs where $j = \lfloor ns \rfloor$ and $k = \lfloor ns + 4 \rfloor$. $\tilde{Y}_{j,k}$ closely approximates $\tilde{U}_{j,k}$ in Fig. 5.2 for both the modified ER and classic ER graphs and the bounds presented in Prop. 13 also appear tight. $\tilde{Y}_{j,k}$'s good approximation of $\tilde{U}_{j,k}$ may be due to $j \approx \mu$ and $k \approx \mu$.

5.4.2 SNAP results

Prop. 12 and Prop. 13 were tested on six real-world graphs from the Stanford Large Network Dataset Collection (SNAP) Leskovec and Krevl [2014]. In all six graphs $|\mathcal{V}_\lambda| = 1$, additional properties of these graphs are listed in Tab. 5.1 where n is the number of vertices, m the number of unordered edges, ρ the edge density, μ the average degree, λ the max degree, and α the assortativity of the graph.

Table 5.1: SNAP graph Leskovec and Krevl [2014] properties.

Graph	n	m	ρ	μ	λ	α
CondMat	23,133	93,497	0.00035	8.1	279	0.13
Enron	36,692	183,831	0.00027	10.0	1,383	-0.11
Facebook	4,039	88,234	0.01082	43.7	1,045	0.06
GrQc	5,242	14,496	0.00106	5.5	81	0.66
HepPh	12,008	118,521	0.00164	19.7	491	0.63
HepTh	9,877	25,998	0.00053	5.3	65	0.27

Table 5.2: Prop. 12 SNAP graph results; Θ_{u_k, y_k} is the relative error.

	Con.	*Enr.*	*Fac.*	*GrQ.*	*HepP.*	*HepT.*
$\frac{1}{\bar{y}_k + \epsilon_k}$	82.6	26.5	3.9	62.4	24.2	141.1
$\frac{1}{\bar{u}_k}$	82.6	26.5	3.9	63.9	24.4	149.7
$\frac{1}{\bar{u}_k}$	83.8	26.5	4.4	71.4	29.0	146.4
$\frac{1}{\bar{y}_k - \epsilon_k}$	85.1	26.5	5.0	83.3	36.3	152.1
Θ_{u_k, y_k}	0.015	0.000	0.131	0.116	0.190	0.022

The results in Tab. 5.2 are for the expected number of star samples needed to find a degree λ vertex. In all cases $1/\bar{y}_k$ is within 19.0% relative error of $1/\bar{u}_k$. It is surprising that the estimator

Table 5.3: Prop. 13 SNAP graph results; $\Theta_{\tilde{U}_{j,k},\tilde{Y}_{j,k}}$ is the relative error.

	Con.	*Enr.*	*Fac.*	*GrQ.*	*HepP.*	*HepT.*
$\frac{1}{\tilde{Y}_{j,k}+\tilde{\epsilon}^U_{j,k}}$	43.9	265.9	175.6	58.6	285.9	41.2
$\frac{1}{\tilde{Y}_{j,k}+\tilde{\epsilon}_{j,k}}$	43.9	265.9	175.6	61.6	285.9	41.2
$\frac{1}{\tilde{U}_{j,k}}$	61.9	265.9	175.6	69.9	285.9	41.2
$\frac{1}{\tilde{Y}_{j,k}}$	51.3	335.3	292.8	65.5	400.7	43.1
$\frac{1}{\tilde{Y}_{j,k}-\tilde{\epsilon}_{j,k}}$	61.9	453.6	880.8	69.9	669.4	45.3
$\frac{1}{\tilde{Y}_{j,k}-\tilde{\epsilon}^U_{j,k}}$	61.9	453.6	880.8	74.2	669.4	45.3
$\Theta_{\tilde{U}_{j,k},\tilde{Y}_{j,k}}$	0.170	0.261	0.668	0.063	0.401	0.048

$1/\bar{y}_k$ of the expected number of star samples needed to find a degree k vertex is close given it was designed for graphs where edges are placed uniformly at random between vertices. Tab. 5.3 gives the corresponding results for the estimator $1/\tilde{Y}_{j,k}$ of the expected number of samples needed to find a degree (j,k) edge in an ER graph. The relative error in these results falls between 4.8% and 66.8%, indicating that for the SNAP graphs $1/\tilde{Y}_{j,k}$ is not as accurate an estimator as $1/\bar{y}_k$.

5.5 Conclusions

This chapter introduced estimators for the expected number of star samples required to find a degree k vertex, $1/\bar{y}_k$, and to find a degree $\{j,k\}$ edge, $1/\tilde{Y}_{j,k}$. To analyze the accuracy of these estimators and specifically the estimator for SS-S on ER graphs, this chapter introduced the modified ER construction and showed that the modified ER graphs $\tilde{G}_\epsilon$ are similar to the classic ER graphs.

Since the introduced estimators are accurate for the former graphs $\tilde{G}_\epsilon$, it is implied that the approximation in given Def. 16 is accurate in both graph types. Finally this chapter shows that the estimators introduced are reasonably accurate in classic and modified ER graphs but are inconsistent on "real-world" SNAP graphs.

5.6 Appendix

Proof of Thm. 10. Objective, show $\mathbb{E}[\tilde{\mathsf{y}}^{(n)}] \to \mathrm{e}^{-s}$, and claim that $\mathrm{Var}(\tilde{\mathsf{y}}^{(n)}) \to 0$, which together imply $\tilde{\mathsf{y}}^{(n)} \xrightarrow{\mathrm{P}} \mathrm{e}^{-s}$.

Proof that $\mathbb{E}[\tilde{\mathsf{y}}^{(n)}] \to \mathrm{e}^{-s}$. Fix $n \in \mathbb{N}$ and $s \in (0,1)$, define the RV $\tilde{\mathsf{m}} \sim \mathrm{Bin}(M,s)$, and set $p_0 = \mathbb{P}(\tilde{\mathsf{m}} = 0) = \bar{s}^M$. Preclude the empty graph, obtained for $\tilde{\mathsf{m}} = 0$, and assume the number of edges is given by the RV $\hat{\mathsf{m}}$ with distribution $\mathbb{P}(\hat{\mathsf{m}} = m) = \mathbb{P}(\tilde{\mathsf{m}} = m)/\bar{p}_0$ for $m \in [M]$. Define the random vector $\tilde{\mathsf{x}} = (\tilde{\mathsf{x}}_1, \ldots, \tilde{\mathsf{x}}_M)$ as holding the occupancy counts of each of the M possible edges

sites. Conditioned on $\hat{\mathsf{m}}$, $\tilde{\mathsf{x}}$ has a Multinomial distribution, i.e., $\tilde{\mathsf{x}}|\hat{\mathsf{m}} \sim \text{Mult}(\hat{\mathsf{m}}, \mathbf{1}_M/M)$, for $\mathbf{1}_M/M$ the M-vector with values $(1/M, \ldots, 1/M)$. By construction, $\sum_{i\in[M]} \tilde{\mathsf{x}}_i = \hat{\mathsf{m}}$. By definition, for any $x = (x_1, \ldots, x_M)$ with $x_1 + \cdots + x_M = \hat{\mathsf{m}}$:

$$\mathbb{P}(\tilde{\mathsf{x}} = x|\hat{\mathsf{m}}) = \binom{\hat{\mathsf{m}}}{x} \prod_{e\in[M]} \left(\frac{1}{M}\right)^{x_i} = \binom{\hat{\mathsf{m}}}{x} \frac{1}{M^{\hat{\mathsf{m}}}}. \tag{5.23}$$

Let the conditional RV $\tilde{\mathsf{x}}_i|\hat{\mathsf{m}} \sim \text{Bin}(\hat{\mathsf{m}}, 1/M)$ denote the occupancy of edge site $i \in [M]$, with conditional mean $\mathbb{E}[\tilde{\mathsf{x}}_i|\hat{\mathsf{m}}] = \hat{\mathsf{m}}\frac{1}{M}$ and conditional variance $\text{Var}(\tilde{\mathsf{x}}_i|\hat{\mathsf{m}}) = \hat{\mathsf{m}}\frac{1}{M}(1 - \frac{1}{M})$. Define the conditional indicator RV $\tilde{\mathsf{y}}_i|\hat{\mathsf{m}} = \mathbf{1}(\tilde{\mathsf{x}}_i = 1)|\hat{\mathsf{m}}$ as indicating if edge site i holds exactly one edge, with:

$$\mathbb{E}[\tilde{\mathsf{y}}_i|\hat{\mathsf{m}}] = \mathbb{P}(\tilde{\mathsf{x}}_i = 1|\hat{\mathsf{m}}) = \hat{\mathsf{m}}\frac{1}{M}\left(1 - \frac{1}{M}\right)^{\hat{\mathsf{m}}-1}. \tag{5.24}$$

The conditional random fraction of non-multiple edges, i.e., the number of the $\hat{\mathsf{m}}$ edges in edge sites with unit occupancy, is then $\tilde{\mathsf{y}}|\hat{\mathsf{m}} = \frac{1}{\hat{\mathsf{m}}}\sum_{i\in[M]} \tilde{\mathsf{y}}_i|\hat{\mathsf{m}}$, with (by linearity of expectation):

$$\mathbb{E}[\tilde{\mathsf{y}}|\hat{\mathsf{m}}] = \frac{1}{\hat{\mathsf{m}}} M\mathbb{E}[\tilde{\mathsf{y}}_i|\hat{\mathsf{m}}] = \left(1 - \frac{1}{M}\right)^{\hat{\mathsf{m}}-1}. \tag{5.25}$$

Define $\hat{p}_0 = 1/\bar{p}_0 = 1/(1-\bar{s}^M)$. The scaled (and unconditioned) expected fraction of non-multiple edges is

$$\begin{aligned}
\left(1 - \frac{1}{M}\right)\frac{\mathbb{E}[\tilde{\mathsf{y}}]}{\hat{p}_0} &= \left(1 - \frac{1}{M}\right)\frac{1}{\hat{p}_0}\mathbb{E}[\mathbb{E}[\tilde{\mathsf{y}}|\hat{\mathsf{m}}]] \\
&= \frac{1}{\hat{p}_0}\sum_{m=1}^{M}\left(1 - \frac{1}{M}\right)^{m}\mathbb{P}(\hat{\mathsf{m}} = m) \\
&= \frac{1}{\hat{p}_0}\sum_{m=1}^{M}\left(1 - \frac{1}{M}\right)^{m}\frac{\mathbb{P}(\tilde{\mathsf{m}} = m)}{\bar{p}_0} \\
&= -p_0 + \sum_{m=0}^{M}\left(1 - \frac{1}{M}\right)^{m}\mathbb{P}(\tilde{\mathsf{m}} = m) \\
&= -p_0 + \mathbb{E}\left[\left(1 - \frac{1}{M}\right)^{\tilde{\mathsf{m}}}\right].
\end{aligned} \tag{5.26}$$

Recall that the PGF $\varphi_{\tilde{\mathsf{m}}}(z) \equiv \mathbb{E}[z^{\tilde{\mathsf{m}}}]$ for $\tilde{\mathsf{m}} \sim \text{Bin}(M, s)$ is $\varphi_{\tilde{\mathsf{m}}}(z) = (\bar{s} + sz)^M$, and as such

$$\mathbb{E}\left[\left(1 - \frac{1}{M}\right)^{\tilde{\mathsf{m}}}\right] = \varphi_{\tilde{\mathsf{m}}}\left(1 - \frac{1}{M}\right) = \left(\bar{s} + s\left(1 - \frac{1}{M}\right)\right)^{M}. \tag{5.27}$$

Combining:

$$\begin{aligned}\mathbb{E}[\tilde{\mathsf{y}}] &= \hat{p}_0\left(1-\frac{1}{M}\right)^{-1}\left(\left(1-\frac{s}{M}\right)^M - p_0\right)\\ &= \frac{1}{1-\bar{s}^M}\left(1-\frac{1}{M}\right)^{-1}\left(\left(1-\frac{s}{M}\right)^M - \bar{s}^M\right).\end{aligned} \tag{5.28}$$

Writing $\tilde{\mathsf{y}} = \tilde{\mathsf{y}}^{(n)}$ to emphasize the dependence upon n, it is immediate from Sec. 5.6 that $\mathbb{E}[\tilde{\mathsf{y}}^{(n)}] \to \mathrm{e}^{-s}$ as $n \uparrow \infty$.

Proof that $\mathrm{Var}(\tilde{\mathsf{y}}^{(n)}) \to 0$. As $\mathbb{E}[\tilde{\mathsf{y}}^{(n)}] \to \mathrm{e}^{-s}$ the variance can be shown approach zero by showing $\mathbb{E}[(\tilde{\mathsf{y}}^{(n)})^2] \to \mathrm{e}^{-2s}$. Observe

$$\begin{aligned}\tilde{\mathsf{y}}^2|\hat{\mathsf{m}} &= \left(\frac{1}{\hat{\mathsf{m}}}\sum_{i\in[M]}\tilde{\mathsf{y}}_i\right)^2\Bigg|\,\hat{\mathsf{m}}\\ &= \frac{1}{\hat{\mathsf{m}}}\left(\frac{1}{\hat{\mathsf{m}}}\sum_{i\in[M]}\tilde{\mathsf{y}}_i^2\right) + \frac{1}{\hat{\mathsf{m}}^2}\sum_{(i,i')\in[M]^2}\tilde{\mathsf{y}}_i\tilde{\mathsf{y}}_{i'}\Bigg|\,\hat{\mathsf{m}}\end{aligned} \tag{5.29}$$

where the second sum is over the $2\binom{M}{2} = M(M-1)$ ordered pairs of distinct edges. As $\tilde{\mathsf{y}}_i^2 = \tilde{\mathsf{y}}_i$, by linearity of expectation:

$$\mathbb{E}[\tilde{\mathsf{y}}^2|\hat{\mathsf{m}}] = \frac{\mathbb{E}[\tilde{\mathsf{y}}|\hat{\mathsf{m}}]}{\hat{\mathsf{m}}} + \frac{M(M-1)}{\hat{\mathsf{m}}^2}\mathbb{P}(\tilde{\mathsf{y}}_i = \tilde{\mathsf{y}}_{i'} = 1|\hat{\mathsf{m}}). \tag{5.30}$$

Observe the conditional event $\{\tilde{\mathsf{y}}_i = \tilde{\mathsf{y}}_{i'} = 1\}|\hat{\mathsf{m}}$ has a Multinomial probability for occupancy $(1, 1, \hat{\mathsf{m}}-2)$, where the three "bins" are edge sites i, i' and $[M]\setminus\{i, i'\}$, with probabilities $1/M, 1/M, 1-2/M$, respectively:

$$\mathbb{P}(\tilde{\mathsf{y}}_i = \tilde{\mathsf{y}}_{i'} = 1|\hat{\mathsf{m}}) = \binom{\hat{\mathsf{m}}}{1,1,\hat{\mathsf{m}}-2}\left(\frac{1}{M}\right)^2\left(1-\frac{2}{M}\right)^{\hat{\mathsf{m}}-2}. \tag{5.31}$$

Substitution and simplification yields

$$\mathbb{E}[\tilde{\mathsf{y}}^2|\hat{\mathsf{m}}] = \frac{1}{\hat{\mathsf{m}}}\left(1-\frac{1}{M}\right)^{\hat{\mathsf{m}}-1} + \left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)^{\hat{\mathsf{m}}-2}\left(1-\frac{1}{\hat{\mathsf{m}}}\right). \tag{5.32}$$

It is immediate to see $\mathbb{E}[\tilde{\mathsf{y}}^2|\hat{\mathsf{m}}] \le g(\hat{\mathsf{m}}, M)$, where:

$$g(m, M) \equiv \frac{1}{m} + \left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)^{-2}\left(1-\frac{2}{M}\right)^m. \tag{5.33}$$

The next objective is an upper bound on the unconditioned quantity $\mathbb{E}[\tilde{\mathsf{y}}^2] = \mathbb{E}_{\hat{\mathsf{m}}}[\mathbb{E}[\tilde{\mathsf{y}}^2|\hat{\mathsf{m}}]] \le$

$\mathbb{E}[g(\hat{\mathsf{m}}, M)]$, where $\mathbb{E}[g(\hat{\mathsf{m}}, M)] =$

$$\mathbb{E}\left[\frac{1}{\hat{\mathsf{m}}}\right] + \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right)^{-2} \mathbb{E}\left[\left(1 - \frac{2}{M}\right)^{\hat{\mathsf{m}}}\right]. \tag{5.34}$$

A similar use of the PGF as above establishes

$$\begin{aligned} \mathbb{E}\left[\left(1 - \frac{2}{M}\right)^{\hat{\mathsf{m}}}\right] &= \frac{1}{\bar{p}_0}\left(\mathbb{E}\left[\left(1 - \frac{2}{M}\right)^{\tilde{\mathsf{m}}}\right] - p_0\right) \\ &= \frac{1}{1 - \bar{s}^M}\left(\left(1 - \frac{2s}{M}\right)^M - \bar{s}^M\right) \end{aligned} \tag{5.35}$$

Substituting Sec. 5.6 and the upper bound on $\mathbb{E}[1/\hat{\mathsf{m}}]$ in Lem. 15 into Eq. (5.34) and taking limits shows that $\lim_{M\to\infty} \mathbb{E}[g(\hat{\mathsf{m}}, M)] = \mathrm{e}^{-2s}$, and as such,

$$\begin{aligned} \lim_{M\to\infty} \mathrm{Var}(\tilde{\mathsf{y}}) &= \lim_{M\to\infty} \mathbb{E}[\tilde{\mathsf{y}}^2] - \lim_{M\to\infty} \mathbb{E}[\tilde{\mathsf{y}}]^2 \\ &\leq \lim_{M\to\infty} \mathbb{E}[g(\hat{\mathsf{m}}, M)] - \lim_{M\to\infty} \mathbb{E}[\tilde{\mathsf{y}}]^2 \\ &\leq \mathrm{e}^{-2s} - \left(\mathrm{e}^{-s}\right)^2 = 0. \end{aligned} \tag{5.36}$$

□

Lemma 15. *Let $\tilde{\mathsf{m}}, \hat{\mathsf{m}}$ be as defined in the proof of Thm. 10, with $p_0 = \mathbb{P}(\tilde{\mathsf{m}} = 0) = \bar{s}^M$. Then*

$$\mathbb{E}\left[\frac{1}{\hat{\mathsf{m}}}\right] \leq \frac{2}{\bar{p}_0}\mathbb{E}\left[\frac{1}{\tilde{\mathsf{m}} + 1}\right] = \frac{2(1 - \bar{s}^{M+1})}{s(M+1)(1 - \bar{s}^M)}. \tag{5.37}$$

Proof. Observe

$$\mathbb{E}\left[\frac{1}{\hat{\mathsf{m}}}\right] = \sum_{m=1}^{M} \frac{\mathbb{P}(\tilde{\mathsf{m}} = m)}{\bar{p}_0 m}, \quad \mathbb{E}\left[\frac{1}{\tilde{\mathsf{m}} + 1}\right] = \sum_{m=0}^{M} \frac{\mathbb{P}(\tilde{\mathsf{m}} = m)}{m+1}. \tag{5.38}$$

Note $1/(2m) \leq 1/(m+1)$ holds for all $m \in [M]$, and so:

$$\frac{\bar{p}_0}{2}\mathbb{E}\left[\frac{1}{\hat{\mathsf{m}}}\right] = \sum_{m=1}^{M} \frac{\mathbb{P}(\tilde{\mathsf{m}} = m)}{2m} \leq \sum_{m=0}^{M} \frac{\mathbb{P}(\tilde{\mathsf{m}} = m)}{m+1} = \mathbb{E}\left[\frac{1}{\tilde{\mathsf{m}} + 1}\right]. \tag{5.39}$$

The result

$$\mathbb{E}\left[\frac{1}{\tilde{\mathsf{m}} + 1}\right] = \frac{1 - \bar{s}^{M+1}}{s(M+1)} \tag{5.40}$$

follows from the absorption identity $\frac{1}{M+1}\binom{M+1}{m+1} = \frac{1}{m+1}\binom{M}{m}$ and standard manipulations. □

The following lemma is used in the simplification of ratio in Thm. 12

Lemma 16. *For any $n \in \mathbb{N}$ and any $n' \in \mathbb{N}$ with $n' \leq n$, define M' as the number of unordered pairs of distinct values from $[n]$ with either or both values in $[n']$. Then*

$$M' = n' \left(n - 1 - \frac{n' - 1}{2} \right). \tag{5.41}$$

Proof. Given a complete graph K_n on $[n]$, where M' is the number of edges incident with one of the nodes in $[n']$. There are $n'(n-1)$ "stubs", as each of the n' nodes in $[n']$ has stubs to $n-1$ other nodes in K_n. Associating each stub with an edge double counts those edges where both endpoints are in $[n']$ of which there are $\binom{n'}{2} = n'(n'-1)/2$ such edges. Subtracting out this term gives the expression for M'. □

Proof of Thm. 12. $U_{j,k}$ is the probability that a selected degree j node has no degree k neighbors under the assumptions that *i*) there are m_j edges with either or both endpoints of degree j, and *ii*) $m_{j,k}$ of these m_j edges have endpoints of degrees j and k. Under the modified ER graph construction, the degree j node has all j of its adjacent edges drawn uniformly at random from the set m_j. It will have no degree k neighbors provided none of those edges are from the set of edges with degrees j and k. As all such j subsets are approximately equally likely, it follows that $U_{j,k} \approx Y_{j,k}$. The word approximate is used because of a key difference between $U_{j,k}$ and $Y_{j,k}$: the former requires that all degree j nodes have j edges, while no such requirement is enforced in $Y_{j,k}$. This section now derives the following Thm. 12 restated below for a $\tilde{G}_\epsilon$ graph.

$$\frac{U_{j,k}}{Y_{j,k}} = \left(\frac{1 - \frac{n_j - 1}{2(n-1)}}{1 - \frac{n_j - 1}{2(n-1-n_k)}} \right)^j \left(\frac{1 - \frac{1}{n_j} \times \frac{1}{1 - \frac{n_j - 1}{2(n-1-n_k)}}}{1 - \frac{1}{n_j} \times \frac{1}{1 - \frac{n_j - 1}{2(n-1)}}} \right)^{m_j - j} \left(\frac{1 - \frac{1}{n_j} \times 1}{1 - \frac{1}{n_j} \times \frac{1 - \frac{n_k}{n-1}}{1 - \frac{n_j - 1}{2(n-1)}}} \right)^{m_{j,k}} \tag{5.42}$$

Each of the three terms in the product is approximately equal to 1 whenever n_k or n_l is small relative to n.

Recall the modified ER construction admits a direct interpretation using balls and bins, with balls corresponding to actual edges and bins corresponding to potential edge sites. There are $M = \binom{n}{2}$ bins and m balls, with the balls placed independently and uniformly at random in the bins. Suppose this has been done and counting reveals the values of $(m_j, m_{j,k})$. Let A_j (with $|A_j|$ denoted by M_j) be the subset of bins for edge sites with one or both endpoints of degree j, the bins labeled WLOG as $[M_j]$. Let B_1 (with $|B_1| = n-1$) denote the subset of bins associated with node 1, these bins labeled WLOG as $[n-1]$. Let $D_{j,k}$ (with $|D_{j,k}|$ denoted by $M_{j,k}$) be the subset of A_j tied to degree k nodes.

Let $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ be the Multinomial random vector of the N bins, with $\sum_{i=1}^{N} \mathsf{x}_i = m$. Define events:

$$I_j : \sum_{i \in A_j} \mathsf{x}_i = m_j,\ I_1 : \sum_{i \in B_1} \mathsf{x}_i = j,\ I_{j,k} : \sum_{i \in D_{j,k}} \mathsf{x}_i = m_{j,k},\ I_{1,\bar{k}} : \sum_{i \in B_1 \cap D_{j,k}} \mathsf{x}_i = 0. \tag{5.43}$$

Observe $B_1 \cap D_{j,k}$ are those bins connecting node 1 with the degree k nodes. The goal is to compute:

$$U_{j,k} = \mathbb{P}(I_{1,\bar{k}} | I_j, I_1, I_{j,k}) = \frac{\mathbb{P}(I_{1,\bar{k}}, I_1 | I_j, I_{j,k})}{\mathbb{P}(I_1 | I_j, I_{j,k})}. \tag{5.44}$$

In words, given that m_j balls are placed in bins A_j, j balls are placed in bins B_1, and $m_{j,k}$ balls are placed in bins $D_{j,k}$, what is the probability that no balls are placed in bins $B_1 \cap D_{j,k}$?

Define three binomial RVs:

- $\mathsf{M}_{j,\bar{k}}$: the number of the $m_j - m_{j,k}$ balls with endpoints j but not k that land in bins tied to vertex 1. In particular, $\mathsf{M}_{j,\bar{k}} \sim \text{Bin}(m_j - m_{j,k}, p_{j,\bar{k}})$, for $p_{j,\bar{k}} = \frac{n-1-n_k}{M_j - M_{j,k}}$.
- $\mathsf{M}_{j,k}$: the number of the $m_{j,k}$ balls with endpoints j and k that land in bins tied to vertex 1. In particular, $\mathsf{M}_{j,k} \sim \text{Bin}(m_{j,k}, p_{j,k})$, for $p_{j,k} = \frac{n_k}{M_{j,k}}$.
- M_1: the number of M_j balls with either or both endpoints of degree j that land in bins tied to vertex 1. In particular, $\mathsf{M}_1 \sim \text{Bin}(m_j, p_1)$, for $p_1 = (n-1)/M_j$.

The first two RVs help compute the numerator and the third the denominator in Eq. (5.44):

- Numerator: the conditioned events $I_j, I_{j,k}$ imply two different independent experiments. First: $m_j - m_{j,k}$ balls for the j but not k edges are placed uniformly at random into the $M_j - M_{j,k}$ bins (each bin for edges of degree j but not of degree k). Second: $m_{j,k}$ balls for the (j,k) edges are placed uniformly at random into $M_{j,k}$ bins (each bin for edges of degrees j and k). Both collections of bins are partitioned into two groups. First, the $M_j - M_{j,k}$ bins tied to edges of degrees j but not of degree k are partitioned into bins tied to vertex 1 and those that are not tied to vertex 1. There are $n-1-n_k$ of the former, and thus $M_j - M_{j,k} - (n-1-n_k)$ of the latter. Second, the $M_{j,k}$ bins tied to edges of degrees j and k are partitioned into bins tied to vertex 1 and those that are not tied to vertex 1. There are n_k of the former and $M_{j,k} - n_k$ of the latter. Note $p_{j,\bar{k}}$ is the probability a ball dropped uniformly at random into the $M_j - M_{j,k}$ bins for edges of degree j but not of degree k lands in a bin tied to node 1, and $p_{j,k}$ is the probability a ball dropped uniformly at random into the $M_{j,k}$ bins for edges of degrees j and

k lands in a bin tied to node 1. It follows that $\mathbb{P}(I_{1,k}, I_1 | I_j, I_{j,k}) = \mathbb{P}(\mathsf{M}_{j,\bar{k}} = j)\mathbb{P}(\mathsf{M}_{j,k} = 0)$, where independence holds by construction.

- Denominator: the event of interest is that there are j balls placed in the $n-1$ bins corresponding to vertex 1, and no distinction is made with respect to the other degrees of those balls. As such I_1 is conditionally independent of $I_{j,k}$ given I_j, i.e., $\mathbb{P}(I_1|I_j, I_{j,k}) = \mathbb{P}(I_1|I_j)$. It follows that $\mathbb{P}(I_1|I_j, I_{j,k}) = \mathbb{P}(\mathsf{M}_1 = j)$.

Substituting this gives

$$\begin{aligned} U_{j,k} &= \frac{\mathbb{P}(\mathsf{M}_{j,\bar{k}} = j)\mathbb{P}(\mathsf{M}_{j,k} = 0)}{\mathbb{P}(\mathsf{M}_1 = j)} \\ &= \frac{\binom{m_j - m_{j,k}}{j} p_{j,\bar{k}}^j (1 - p_{j,\bar{k}})^{m_j - m_{j,k} - j} (1 - p_{j,k})^{m_{j,k}}}{\binom{m_j}{j} p_1^j (1 - p_1)^{m_j - j}} \\ &= \frac{\binom{m_j - m_{j,k}}{j}}{\binom{m_j}{j}} \left(\frac{p_{j,\bar{k}}}{p_1}\right)^j \left(\frac{1 - p_{j,\bar{k}}}{1 - p_1}\right)^{m_j - j} \left(\frac{1 - p_{j,k}}{1 - p_{j,\bar{k}}}\right)^{m_{j,k}} \end{aligned} \tag{5.45}$$

It remains to derive the expressions in Eq. (5.42).

$$\begin{aligned} \frac{p_{j,\bar{k}}}{p_1} &= \frac{\frac{n-1-n_k}{M_j - M_{j,k}}}{\frac{n-1}{M_j}} = \frac{M_j}{M_j - M_{j,k}} \frac{n - 1 - n_k}{n - 1} \\ \frac{1 - p_{j,\bar{k}}}{1 - p_1} &= \frac{1 - \frac{n-1-n_k}{M_j - M_{j,k}}}{1 - \frac{n-1}{M_j}} = \frac{M_j}{M_j - M_{j,k}} \frac{(M_j - (n-1)) - (M_{j,k} - n_k)}{M_j - (n-1)} \\ \frac{1 - p_{j,k}}{1 - p_{j,\bar{k}}} &= \frac{1 - \frac{n_k}{M_{j,k}}}{1 - \frac{n-1-n_k}{M_j - M_{j,k}}} = \frac{M_j - M_{j,k}}{(M_j - M_{j,k}) - (n - 1 - n_k)} \frac{M_{j,k} - n_k}{M_{j,k}} \end{aligned} \tag{5.46}$$

Using $M_{j,k} = n_j n_k$, Lem. 16, and extensive algebra yield the given expressions. □

Chapter 6: Star sampling with and without replacement

6.1 Introduction

Consider a simple undirected graph $G = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Now consider some vertex property, defined via a function $f : \mathcal{V} \to \mathcal{P}$, for $\mathcal{P}$ the set of property values, where f may or may not depend upon G. Then $\mathcal{V}^* \equiv f^{-1}(\mathcal{P}^*)$, for $\mathcal{P}^* \subset \mathcal{P}$, is the subset of vertices holding property values of interest (i.e., in $\mathcal{P}^*$). Consider the problem of finding any vertex in $\mathcal{V}^*$. This chapter evaluates the performance of three related random sampling approaches to this problem, called *star sampling*, described in detail below, that differ in terms of which part of the sample is replaced.

Star sampling. Using random sampling to search for a vertex of interest is often suitable for large and/or dynamic graphs, where either the order/size and/or the rapid evolution of the graph precludes holding the graph in local memory. In such cases the searcher may be required to *query* the graph, requesting the property value of either a *random* vertex (as in sampling) or a *particular* vertex (as in guided search). *Star sampling* (SS) is a variant on vertex sampling in which each sample returns not only the property value $f(v)$ of the selected vertex $v \in \mathcal{V}$, termed the *star center*, but also the property values $f(\Gamma_v)$ of its one-hop neighbors Γ_v (where $\Gamma_v \equiv \{u \in \mathcal{V} : uv \in \mathcal{E}\}$ is the neighborhood of v, and $uv \in \mathcal{E}$ denotes vertices u and v are joined by an edge), henceforth termed the *star endpoints*. Star sampling is the simple idea of selecting a vertex at random, and then checking whether either that vertex or any of its neighbors holds the property of interest.

Cost. Although large graphs are encoded in a variety of ways, it is often the case that the data structure corresponding to each vertex (e.g., the profile of a particular member in a social network) includes the list of neighbors of that vertex (e.g., the social connections of that member in the network). Star sampling is a practical sampling paradigm whenever such neighbor information is available. That said, the property of interest for the neighbors of a given vertex, i.e., the values $f(\Gamma_v)$, may or may not be readily available. The property may be readily available if it is easily computable, or if it is not easily computable, but has been *precomputed* and stored in the data structure for the vertex. To address this issue this chapter considers two natural cost models (*unit* and *linear*), where cost is measured either as the number of star samples (unit cost) or as the number of vertices (linear cost) for which property values are queried / computed. Unit cost is most natural for the case when the primary cost incurred is the query of the star itself and property values of

neighbors are readily available, while linear cost is most natural for the case where the primary cost incurred is the computation of the property value.

Scenarios. This chapter considers two distinct search scenarios in which the property of interest $\mathcal{P}^*$ is *i*) related to the graph G, or *ii*) independent of G. The performance of Scenario *i*) is not possible to analyze in general, as many possible graph-related properties are possible. As such, this chapter considers an important representative scenario, namely, where the property of interest is having the maximum degree (λ) of the graph, i.e., $\mathcal{V}^*$ is the set of vertices with degree λ and λ is known *a priori*. In a social network, this corresponds to the problem of identifying any individual with the largest number of contacts, when that largest number is known or can be estimated. Scenario *ii*) is studied in general, as the cardinality of the target set $|\mathcal{V}^*|$ is a sufficient statistic for $\mathcal{V}^*$, i.e., all target sets $\mathcal{V}^*$ of the same cardinality $|\mathcal{V}^*|$, chosen independently of the graph, share a common difficulty to find a member of $\mathcal{V}^*$.

Variants. Three SS variants are considered:

- *SS with replacement (SS-R)*: the star center is selected uniformly at random from the set of vertices;
- *SS without center replacement (SS-C)*: the star center is selected uniformly at random from the set of *remaining* vertices; the star center (along with its adjacent edges) is removed from the graph after the query;
- *SS without star replacement (SS-S)*: the star center is selected uniformly at random from the set of *remaining* vertices; the entire star (center, endpoints, and all adjacent edges) is removed from the graph after the query.

Urn sampling. The motivation in considering these variants is to understand their relative performance, in a manner similar to the elementary case of sampling balls from an urn. When seeking any one of n^* marked balls out of a total of $n \geq n^*$ balls in an urn, sampling *with* replacement requires on average n/n^* samples; this follows immediately from the observation that the number of draws until the first success, say c^{R}, is a geometric random variable (RV) with success probability n^*/n, and expectation $\mathbb{E}[\mathsf{c}^{\mathrm{R}}] = n/n^*$. In contrast, sampling *without* replacement requires a random number of draws, c^{NR}, with expectation $\mathbb{E}[\mathsf{c}^{\mathrm{NR}}] = (n+1)/(n^*+1)$ (c.f. Prop. 20 in Sec. 6.10). Thus, the performance ratio of the expected number of samples with vs. without replacement is $\mathbb{E}[\mathsf{c}^{\mathrm{R}}]/\mathbb{E}[\mathsf{c}^{\mathrm{NR}}] = (1+1/n^*)/(1+1/n)$. For $n^* \ll n$, sampling without replacement improves the average search time by at most a factor of two, relative to sampling with replacement, with equality for $n^* = 1$.

Contributions. The three primary contributions are:

- Exact and approximate expected unit and linear cost expressions (for ER random graphs) for the three star sampling variants, and comparisons with simulation results suggesting the approximations are quite accurate;
- Numerical evidence (from ER random graphs) that the approximate expected unit cost of the three SS variants in the unit cost model are more or less identical, while the approximate expected linear costs are notably distinct;
- Numerical evidence that the approximate expected unit and linear cost expressions are surprisingly accurate for "real-world", i.e., non-ER, graphs.

Outline. The chapter is organized as follows. Sec. 6.2 provides basic notation and definitions. Sec. 6.3 studies performance under the *unit* cost model. Sec. 6.4 shows there is *no ordering* on the performance of the three star sampling variants under the unit cost model that holds for all graphs and target sets. Sec. 6.5 studies performance under the *linear* cost model. Sec. 6.6 gives numerical results for the performance estimates given in this chapter on "real-world" graphs. Sec. 6.7 discusses related work and Sec. 6.8 gives brief conclusions. Several appendices hold supporting results and longer proofs.

6.2 Notation, Sampling Model, Background

6.2.1 Notation

Let $a \equiv b$ denote a and b are equal by definition. Let $[n]$ denote $\{1, \ldots, n\}$, for $n \in \mathbb{N}$. Scalar random variables are denoted in a lowercase sans-serif font, e.g., x, k, in this chapter specifically graph and set-valued random variables are denoted with uppercase sans-serif font, e.g., $\mathsf{G}, \mathsf{V}, \mathsf{E}$. Expectation is denoted $\mathbb{E}[\cdot]$, and probability is denoted $\mathbb{P}(\cdot)$. If $\mathcal{U}$ is a set then $\mathsf{u} \sim \mathrm{Uni}(\mathcal{U})$ denotes a member of $\mathcal{U}$ selected uniformly at random. This chapter uses the following graph and sampling notation:

- *Order, size, edges.* An undirected and simple graph of order n is denoted $G = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V} \equiv [n]$ and edge set $\mathcal{E}$; size is denoted by $m \equiv |\mathcal{E}|$. An undirected edge is denoted uv.
- *Neighborhoods.* Let $\Gamma_v \equiv \{u \in \mathcal{V} : uv \in \mathcal{E}\}$ denote the (direct) neighbors of v, $\Gamma_v^e \equiv \Gamma_v \cup \{v\}$ the *extended* neighborhood of v, $\mathcal{N}_v \equiv \{uv \in \mathcal{E}\}$ the edge neighborhood of v, i.e., the edges adjacent to v, and

$$\mathcal{N}_v^e \equiv \bigcup_{u \in \Gamma_v} \mathcal{N}_u, \tag{6.1}$$

the *extended* edge neighborhood of v, i.e., all edges adjacent to v or any of v's neighbors. Observe Γ_v^e is a star sample with star center v and star endpoints Γ_v. For $\mathcal{V}^* \subseteq \mathcal{V}$, let $\Gamma(\mathcal{V}^*) \equiv (\bigcup_{v\in\mathcal{V}^*} \Gamma_v) \setminus \mathcal{V}^*$ denote neighbors of $\mathcal{V}^*$ not including $\mathcal{V}^*$, and let $\mathcal{V}_G^{e,*} \equiv \Gamma^e(\mathcal{V}^*) \equiv \bigcup_{v\in\mathcal{V}^*} \Gamma_v^e$ denote $\mathcal{V}^*$ and its neighbors.

- *Degrees.* Let $d_v \equiv |\Gamma_v|$ denote the degree of v, and $d_v^e \equiv |\Gamma_v^e|$ the "extended degree", i.e., $d_v^e = d_v + 1$. Let $\mathcal{D} \equiv \bigcup_{v\in\mathcal{V}} d_v$ denote the set of degrees found in G, $\lambda \equiv \max(\mathcal{D})$ the maximum degree, and $\mathcal{V}_\lambda$ the vertices with maximum degree. Partition $\mathcal{V}$ by degree into subsets $(\mathcal{V}_G(k), k \in \mathcal{D})$, with $\mathcal{V}_G(k) \equiv \{v \in \mathcal{V} : d_v = k\}$ the set of vertices with degree k and $w_G(k) \equiv |\mathcal{V}_G(k)|/n$ the fraction of vertices with degree k, to obtain the degree *distribution* of G, denoted $\mathbf{w}_G \equiv (w_G(k), k \in \mathcal{D})$ (with $\sum_{k\in\mathcal{D}} w_G(k) = 1$). The *expected degree* of a randomly selected vertex is

$$\mu \equiv \mathbb{E}[d_{\mathsf{v}}] = \sum_{k\in\mathcal{D}} k w_G(k) = \frac{1}{n}\sum_{v\in\mathcal{V}} d_v. \tag{6.2}$$

6.2.2 Erdős Rényi Graph Properties

An Erdős-Rényi (ER) random graph $\mathsf{G} = (\mathcal{V}, \mathsf{E})$ has parameters (n, s), where $n \in \mathbb{N}$ denotes the order, $\mathcal{V} = [n]$, and $s \in (0, 1)$ denotes the edge probability. A realization of an ER random graph, i.e., of the random edge set E, is obtained by adding each of the $\binom{n}{2}$ possible edges independently at random with probability s. The random size of G, denoted $\mathsf{m} \equiv |\mathsf{E}|$, is a binomial RV, i.e., $\mathsf{m} \sim \text{Bin}\left(\binom{n}{2}, s\right)$, as is the degree $d_{\mathsf{G}}(\mathsf{v})$ of a randomly selected vertex $\mathsf{v} \sim \text{Uni}(\mathcal{V})$, namely, $d_{\mathsf{G}}(\mathsf{v}) \sim \text{Bin}(n-1, s)$. The RV $\mathsf{s} \equiv \mathsf{m}/\binom{n}{2}$ is the *edge density*, and the RV $\mathsf{d} \equiv (n-1)\mathsf{s} = 2\mathsf{m}/n$ is the *average degree* of G. Given an ER graph G, the RV d is the expected degree of a vertex selected uniformly at random.

This chapter focuses its analysis of star sampling on ER graphs since they are closed under both SS-C and SS-S, i.e., if either a star-center (SS-C) or a star (SS-S) is removed from an ER graph the resulting graph is still an ER graph, albeit one with a different (reduced) order. This follows intuitively from the fact that the presence or absence of an edge is independent across edges, but is established formally in Lem. 18 and Lem. 19 in Sec. 6.9.

6.2.3 Sampling Model

Random star sampling from the given graph $G = (\mathcal{V}, \mathcal{E})$ produces a sequence of random graphs, denoted $(\mathsf{G}_t, t \in \mathbb{N})$, where $\mathsf{G}_t = (\mathsf{V}_t, \mathsf{E}_t)$ is the random graph after sample t. Both the vertex set and edge set are (in general) random variables. It is convenient to denote the given graph as $G_0 = (\mathcal{V}_0, \mathcal{E}_0)$, i.e., the initial member of the list of random graphs, corresponding to $t = 0$. The

target set is denoted by $\mathcal{V}^*$ or $\mathcal{V}_0^*$. Define the sequence of random vertex sets $(\mathsf{V}_t^*, t \in \mathbb{N})$, where $\mathsf{V}_t^* \equiv \mathsf{V}_t \cap \mathcal{V}_0^*$ holds the members of the initial target set still "alive" after t samples. The sampling construction ensures the random sets are nested: $\mathsf{V}_{t+1} \subseteq \mathsf{V}_t$, $\mathsf{E}_{t+1} \subseteq \mathsf{E}_t$, and $\mathsf{V}_{t+1}^* \subseteq \mathsf{V}_t^*$. Recalling Sec. 6.1, the three star sampling variants are as follows.

Definition 20. *SS with replacement (SS-R). Generate the IID random sequence of star centers* $(\mathsf{v}_t, t \in \mathbb{N})$*, with* $\mathsf{v}_t \sim \mathrm{Uni}(\mathcal{V}_0)$*. As SS-R uses replacement,* $\mathsf{G}_t = G_0$ *for all* t*.*

Definition 21. *SS without center replacement (SS-C). Generate the random sequence of star centers* $(\mathsf{v}_t, t \in [n])$*, with* $\mathsf{v}_t \sim \mathrm{Uni}(\mathsf{V}_{t-1})$*, and update the graph by removing the star center, i.e.,* $\mathsf{V}_t = \mathsf{V}_{t-1} \setminus \{\mathsf{v}_t\}$*, and the edges in the edge neighborhood of the star center, i.e.,* $\mathsf{E}_t = \mathsf{E}_{t-1} \setminus \mathcal{N}_{\mathsf{v}_t}$*.*

Definition 22. *SS without star replacement (SS-S). Generate the random sequence of star centers* $(\mathsf{v}_t, t \in [n])$*, with* $\mathsf{v}_t \sim \mathrm{Uni}(\mathsf{V}_{t-1})$*, and update the graph by removing the star, i.e.,* $\mathsf{V}_t = \mathsf{V}_{t-1} \setminus \Gamma^e_{\mathsf{G}_{t-1}}(\mathsf{v}_t)$*, and the edges in the* extended *edge neighborhood of the star center, i.e.,* $\mathsf{E}_t = \mathsf{E}_{t-1} \setminus \mathcal{N}^e_{\mathsf{v}_t}$*.*

As a brief aside, Lem. 17 gives the expected number of edges removed from an ER graph. Consider a star sample with star center $\mathsf{v} \sim \mathrm{Uni}(\mathcal{V})$ of an ER graph G with parameters (n, s); recall v has degree distribution $\mathsf{d} \sim \mathrm{Bin}(n-1, s)$.

Lemma 17. *Given an ER random graph* G *with parameters* (n, s)*, a randomly selected vertex* v *as star center, and conditioned on the degree* d *of* v*, the random number of edges in the extended edge neighborhood of* v*, denoted* $\mathsf{g} \equiv |\Gamma^e_{\mathsf{G}}(\mathsf{v})|$*, has a binomial distribution*

$$\mathsf{g}|\mathsf{d} \sim \mathsf{d} + \mathrm{Bin}\left(\binom{\mathsf{d}}{2} + \mathsf{d}(n - \mathsf{d} - 1), s\right), \tag{6.3}$$

with (unconditional) expectation

$$\mathbb{E}[\mathsf{g}] = (n-1)s(1 + (n/2 - 1)(2 - s)s). \tag{6.4}$$

The asymptotic (in n*) ratio of* $\mathbb{E}[\mathsf{g}]$ *to the expected total number of edges in the graph* $\binom{n}{2}s$ *in the graph is* $\lim_{n\to\infty} \mathbb{E}[\mathsf{g}]/(\binom{n}{2}s) = (2 - s)s$*.*

Proof. A star sample has two "types" of vertices (the star center v and its d neighbors) and three "types" of edges, namely, *i*) "neighbor" edges connecting v with Γ_{v}, *ii*) "internal" edges with both endpoints in Γ_{v}, and *iii*) "external" edges with one endpoint in Γ_{v} and the other in $\mathcal{V} \setminus \Gamma^e_{\mathsf{v}}$. There are d neighbor edges by assumption, $\binom{\mathsf{d}}{2}$ potential internal edges, and $\mathsf{d}(n - \mathsf{d} - 1)$ potential external

edges, where (due to the ER random graph properties) all potential edges are present or absent independently with probability s. This explains Eq. (6.3).

As $\mathsf{d} \sim \text{Bin}(n-1, s)$, it follows that $\mathbb{E}[\mathsf{d}] = (n-1)s$ and $\mathbb{E}[\mathsf{d}^2] = (n-1)s + 2\binom{n-1}{2}s^2$, and with these Eq. (6.4) is obtained by conditional expectation and simple algebra:

$$\begin{aligned}
\mathbb{E}[\mathsf{g}] &= \mathbb{E}[\mathbb{E}[\mathsf{g}|\mathsf{d}]] \\
&= \mathbb{E}\left[\mathsf{d} + \left(\binom{\mathsf{d}}{2} + \mathsf{d}(n-\mathsf{d}-1)\right)s\right] \\
&= \mathbb{E}\left[\left(1 + \left(n - \frac{3}{2}\right)s\right)\mathsf{d} - \frac{1}{2}s\mathsf{d}^2\right] \\
&= \left(1 + \left(n - \frac{3}{2}\right)s\right)\mathbb{E}[\mathsf{d}] - \frac{1}{2}s\mathbb{E}[\mathsf{d}^2] \qquad (6.5)
\end{aligned}$$

The limit of $\mathbb{E}[\mathsf{g}]/(\binom{n}{2}s)$ as $n \uparrow \infty$ follows from Eq. (6.4). □

Observe $\mathbb{E}[\mathsf{g}]$ is the average number of edges removed from a star sample under SS-S.

The unit and linear costs for SS are defined below.

Definition 23. *The* unit cost *of a SS is the random number of samples until a star, either the star center or one of the star endpoints, intersects the target set $\mathcal{V}^*$, i.e.,*

$$\mathsf{c}_u(G, \mathcal{V}^*) \equiv \min\{t \in \mathbb{N} : \Gamma^e_{\mathsf{G}_{t-1}}(\mathsf{v}_t) \cap \mathcal{V}^* \neq \emptyset\}. \qquad (6.6)$$

This cost is a function of the sequence of random graphs $(\mathsf{G}_t, t \in \mathbb{N})$. The expected unit cost *of a SS, denoted*

$$c_u(G, \mathcal{V}^*) \equiv \mathbb{E}[\mathsf{c}_u(G, \mathcal{V}^*)], \qquad (6.7)$$

is the expectation of $\mathsf{c}_u(G, \mathcal{V}^*)$, taken with respect to the distribution over all possible realizations of graph sequences induced by SS that begin with $G_0 = G$.

Definition 24. *The* linear cost *of a SS is the random sum of the extended degrees of the randomly selected vertex centers from each star sample until a star, either the star center or one of the star endpoints, intersects the target set $\mathcal{V}^*$, i.e.,*

$$\mathsf{c}_l(G, \mathcal{V}^*) \equiv \sum_{t=1}^{\mathsf{c}_u(G,\mathcal{V}^*)} d^e_{\mathsf{v}_t}. \qquad (6.8)$$

This cost is a function of the sequence of random graphs $(\mathsf{G}_t, t \in \mathbb{N})$. The expected linear cost *of a*

SS, denoted

$$c_l(G, \mathcal{V}^*) \equiv \mathbb{E}[\mathsf{c}_l(G, \mathcal{V}^*)], \tag{6.9}$$

is the expectation of $\mathsf{c}_l(G, \mathcal{V}^*)$*, taken with respect to the distribution over all possible realizations of graph sequences induced by SS that begin with* $G_0 = G$*.*

Fig. 6.1 shows two possible data structures for storing a (large) graph: the first (second) naturally corresponds with the unit (linear) cost model. Under the unit cost model each vertex is represented by four entries: its id, property value, neighbors, and neighbor property values. Under the linear cost model each vertex is represented by the first three of these entries, but the neighbor property values are not available. The memory required under unit cost is order $2n + 2\lambda n$ and under linear cost is order $2n + \lambda n$, i.e., an added cost of order λn.

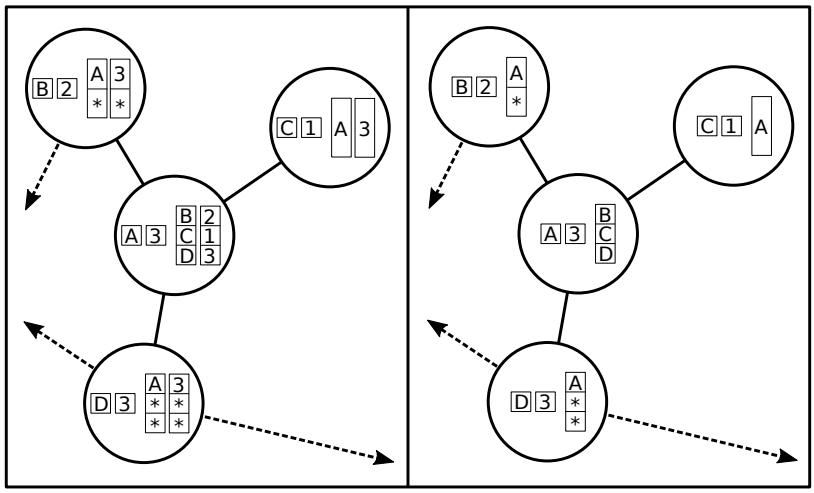

Figure 6.1: Graph representation when property is vertex degree. *Left:* unit cost model (id, prop. value, neighbors, neighbor degree, neighbor prop. value); *right:* linear cost model (id, prop. value, neighbors, neighbor degree).

6.2.4 Taylor series approximation of the expectation of a ratio

Let (y, z) be a pair of continuous RVs, with $\mathsf{z} \neq 0$ almost surely, means $(\mu_\mathsf{y}, \mu_\mathsf{z})$, variances $(\sigma^2_\mathsf{y}, \sigma^2_\mathsf{z})$, and covariance $\mathrm{Cov}(\mathsf{y}, \mathsf{z})$. Consider the expectation of their ratio, i.e., $\mathbb{E}\left[\frac{\mathsf{y}}{\mathsf{z}}\right]$. The following result is found in Van Kempen and Van Vliet [2000] and Seltman [2017].

Proposition 14 (Van Kempen and Van Vliet [2000], Seltman [2017])**.** *The first and second order*

Taylor series approximations of $\mathbb{E}\left[\frac{\mathsf{y}}{\mathsf{z}}\right]$ *around* $(\mu_{\mathsf{y}}, \mu_{\mathsf{z}})$ *are*

$$\mathbb{E}\left[\frac{\mathsf{y}}{\mathsf{z}}\right] = \frac{\mu_{\mathsf{y}}}{\mu_{\mathsf{z}}} + a_1 \tag{6.10}$$

$$\mathbb{E}\left[\frac{\mathsf{y}}{\mathsf{z}}\right] = \frac{\mu_{\mathsf{y}}}{\mu_{\mathsf{z}}} + \frac{\sigma_{\mathsf{z}}^2 \mu_{\mathsf{y}}}{\mu_{\mathsf{z}}^3} - \frac{\mathrm{Cov}(\mathsf{y},\mathsf{z})}{\mu_{\mathsf{z}}^2} + a_2. \tag{6.11}$$

where (a_1, a_2) *are "small" remainder terms.*

In particular, the error associated with a first-order Taylor series approximation is, to second order,

$$\epsilon_1 \equiv \left| \mathbb{E}\left[\frac{\mathsf{y}}{\mathsf{z}}\right] - \frac{\mu_{\mathsf{y}}}{\mu_{\mathsf{z}}} \right| \leq \left| \frac{\sigma_{\mathsf{z}}^2 \mu_{\mathsf{y}}}{\mu_{\mathsf{z}}^3} - \frac{\mathrm{Cov}(\mathsf{y},\mathsf{z})}{\mu_{\mathsf{z}}^2} \right|. \tag{6.12}$$

6.3 Unit cost model

The expected unit costs of SS-R and SS-C are given in Sec. 6.3.1, and the expected unit cost of SS-S is given in Sec. 6.3.2. The (exact) results in Sec. 6.3.1 are given both for an arbitrary graph G and in expectation over the class of ER random graphs, while the (approximate) results in Sec. 6.3.2 are only given in expectation over the class of ER random graphs.

6.3.1 SS-R and SS-C

Let $G = (\mathcal{V}, \mathcal{E})$ be an arbitrary graph of order n, and let $\mathcal{V}^* \subseteq \mathcal{V}$ be an arbitrary target set. In particular, $\mathcal{V}^*$ may or may not depend upon G, as described in Sec. 6.1. The extended neighborhood $\Gamma_G^e(\mathcal{V}^*)$ contains the target set $\mathcal{V}^*$ and its neighbors in G. Observe the equivalence: a star sample Γ_v^e intersects $\mathcal{V}^*$ if and only iff $v \in \Gamma^e(\mathcal{V}^*)$. This observation yields the expected unit cost of SS-R and SS-C. Set $n_G^{e,*} \equiv |\Gamma_G^e(\mathcal{V}^*)|$.

Fact 1 (Unit cost of SS-R)**.** *Under SS-R, for any graph* G *and any target set* $\mathcal{V}^*$*, the unit cost* c_u *in Def. 23 is a geometric RV with success probability* $n_G^{e,*}/n$*, i.e.,* $\mathsf{c}_u^{\mathrm{SS-R}} \sim \mathrm{geo}(n_G^{e,*}/n)$*, and the expected unit cost is* $c_u^{\mathrm{SS-R}} = n/n_G^{e,*}$*.*

Proof. Unit cost of SS-R is the random number of independent Bernoulli trials until the first "success", i.e., the random star intersects the target set, or equivalently, the random star center intersects the extended neighborhood of the target set. □

Fact 2 (Unit cost of SS-C)**.** *Under SS-C, for any graph* G *and any target set* $\mathcal{V}^*$*, the expected unit cost* $c_u^{\mathrm{SS-C}}$ *in Def. 23 is* $c_u^{\mathrm{SS-C}} = (n+1)/(n_G^{e,*}+1)$*.*

Proof. SS-C with target set $\mathcal{V}^*$ on a graph G with extended neighborhood $\Gamma^e(\mathcal{V}^*)$ is equivalent to sampling without replacement from an urn with n balls of which $n_G^{e,*}$ are marked until a marked ball is drawn. The expected number of samples is $(n+1)/(n_G^{e,*}+1)$ (c.f. Prop. 20 in Sec. 6.10). □

Remark 3 (SS-C outperforms SS-R)**.** *Fact 1 and Fact 2 show the expected number of SS-R samples exceeds the expected number of SS-C samples:* $c_u^{\mathrm{SS-R}} > c_u^{\mathrm{SS-C}}$*, but this improvement is negligible for* $n^* \ll n$*, i.e.,* $\lim_{n\to\infty} c_u^{\mathrm{SS-R}}/c_u^{\mathrm{SS-C}} = 1$.

Next the two previous facts are adapted to the case where the initial graph is an ER random graph $\mathsf{G}_0 = (\mathcal{V}_0, \mathsf{E}_0)$ with parameters (n, s), and where the expectation is with respect to both the graph and sampling distributions. Define the RVs:

- $\mathsf{n}_t \equiv |\mathsf{V}_t|$: order of graph G_t (note $n_0 = n$);
- $\mathsf{n}_t^* \equiv |\mathsf{V}_t^*|$: number of vertices from the target set in G_t;
- $\mathsf{n}_t^{e,*} \equiv |\mathsf{V}_t^{e,*}|$ (where $\mathsf{V}_t^{e,*} \equiv \Gamma_{\mathsf{G}_t}^e(\mathsf{V}_t^*)$): number of vertices in extended neighborhood of target set in G_t.

Recall the two search scenarios described in Sec. 6.1:

i) V_0^* is the set of maximum degree vertices in G_0. In this case $\mathsf{n}_0^* = |\mathsf{V}_0^*|$, and $\mathsf{V}_0^{e,*}$ holds all vertices that either have max degree or are adjacent to a max degree vertex. As shown in Chap. 3, the expected number of max degree vertices is close to one for ER random graphs where n is large and s is "small". As this is the scenario of practical importance, this chapter assumes this to be the case, and approximate $\mathsf{n}_0^* \approx 1$. As such, $\mathsf{n}_0^{e,*} \approx 1 + \mathsf{d}^{\max}$, i.e., the (unique) random maximum degree vertex and its neighbors. As shown in Chap. 8,

$$\begin{aligned} \mathbb{E}[\mathsf{d}^{\max}] &\approx ns + \sqrt{ns\bar{s}}\tilde{\mu}(n) \\ \mathrm{Var}(\mathsf{d}^{\max}) &\approx ns\bar{s}\frac{\pi^2}{12\log n}. \end{aligned} \tag{6.13}$$

where

$$\tilde{\mu}(n) \equiv \sqrt{2\log n} - \frac{\log(4\pi\log n)}{2\sqrt{2\log n}} + \frac{\gamma}{\sqrt{2\log n}}, \tag{6.14}$$

and $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

ii) $\mathcal{V}_0^*$ is independent of G_0. Observe all such sets with the same $n_0^* = |\mathcal{V}_0^*|$ are of equivalent difficulty for search. Recalling $\mathsf{n}_0^{e,*}$ as the random order of the extended neighborhood of the

target set in G_0, it is evident, by construction, that

$$\mathsf{n}_0^{e,*} \sim n_0^* + \mathrm{Bin}(n - n_0^*, 1 - \bar{s}^{n_0^*}). \tag{6.15}$$

This holds as each vertex v in $\mathcal{V}_0 \setminus \mathcal{V}_0^*$ is connected to $\mathcal{V}_0^*$ (independently of other vertices) if there exists an edge (or edges) from v to $\mathcal{V}_0^*$, which happens with probability $1 - \bar{s}^{n_0^*}$. In particular,

$$\begin{aligned} \mathbb{E}[\mathsf{n}_0^{e,*}] &= n_0^* + (n - n_0^*)(1 - \bar{s}^{n_0^*}) \\ \mathrm{Var}(\mathsf{n}_0^{e,*}) &= (n - n_0^*)\bar{s}^{n_0^*}(1 - \bar{s}^{n_0^*}). \end{aligned} \tag{6.16}$$

Numerical results for the estimates of $\mathbb{E}[\mathsf{d}^{\max}]$ in Eq. (6.13) and $\mathbb{E}[\mathsf{n}_0^{e,*}]$ in Eq. (6.16) are shown in Fig. 6.2. The top plot in Fig. 6.2 gives the expected size of extended target set $\mathbb{E}[\mathsf{n}_0^{e,*}]$ as a function of the edge density s under Scenarios i) and ii). The bottom plot gives the difference between the true value of $\mathbb{E}[\mathsf{n}_0^{e,*}]$ and its estimate. In Scenario i) the estimate of $\mathbb{E}[\mathsf{n}_0^{e,*}]$ is below the empirical value, which may be due to the de Moivre-Laplace approximation used in Chap. 8 deriving Eq. (6.13). In Scenario ii) the estimate of $\mathbb{E}[\mathsf{n}_0^{e,*}]$ appears quite accurate.

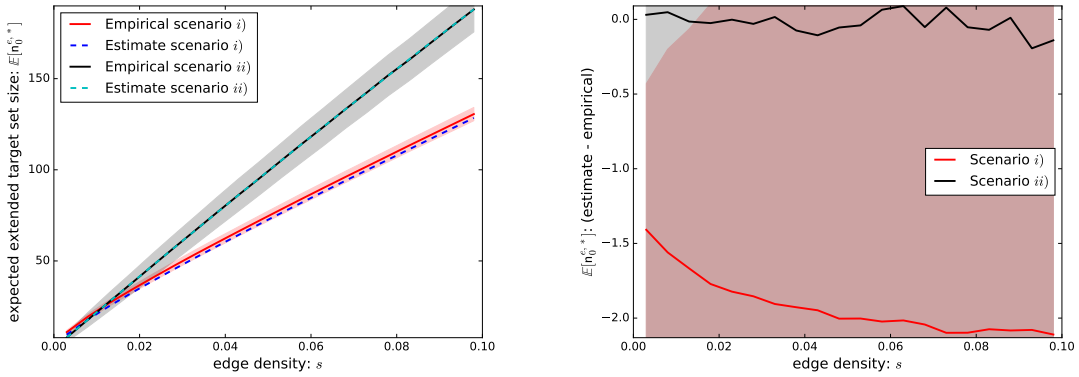

Figure 6.2: Expected extended target set $\mathbb{E}[\mathsf{n}_0^{e,*}]$ (top) and its approximation error (bottom): $n = 1000$, 10k graphs, shaded regions show empirical standard deviation.

The next result, from Lew [1976], is leveraged in Prop. 15 below.

Fact 3 (Bounds on inverse moments, Lew [1976] (Eq. 3.1, p. 729))**.** *Let a random variable* x *have mean* μ*, variance* σ^2*, and minimum support* $x_{\min} > 0$ *(i.e.,* $\mathbb{P}(\mathsf{x} \geq x_{\min}) = 1$*). Then*

$$\frac{1}{\mu} \leq \mathbb{E}\left[\frac{1}{\mathsf{x}}\right] \leq \frac{1}{x_{\min}} \frac{\sigma^2 + (\mu - x_{\min})x_{\min}}{\sigma^2 + (\mu - x_{\min})\mu}. \tag{6.17}$$

Specializing the above result to $\mathsf{x} \sim \mathrm{Bin}(m,p)$ *a binomial RV, let* $a > 0$*. Then*

$$\frac{1}{a+mp} \leq \mathbb{E}\left[\frac{1}{a+\mathsf{x}}\right] \leq \frac{a+1-p}{a(a+1+(m-1)p)}. \tag{6.18}$$

Proposition 15 (Unit cost of SS-R and SS-C for ER graph)**.** *Fix the initial graph as an ER random graph* G_0 *with parameters* (n,s)*, and the extended target set cardinality* $\mathsf{n}_0^{e,*}$*. The expected unit cost under SS-R and SS-C has bounds*

$$\underline{c}_u^{\mathrm{SS-R}} \leq c_u^{\mathrm{SS-R}} \leq \bar{c}_u^{\mathrm{SS-R}}, \quad \underline{c}_u^{\mathrm{SS-C}} \leq c_u^{\mathrm{SS-C}} \leq \bar{c}_u^{\mathrm{SS-C}} \tag{6.19}$$

Under Scenario i), where V_0^* *is the set of maximum degree vertices, and* $\mathbb{E}[\mathsf{d}^{\max}]$ *and* $\mathrm{Var}(\mathsf{d}^{\max})$ *are given in Eq.* (6.13)*:*

$$\begin{aligned}
\underline{c}_u^{\mathrm{SS-R}} &\equiv \frac{n}{1+\mathbb{E}[\mathsf{d}^{\max}]} \\
\bar{c}_u^{\mathrm{SS-R}} &\equiv \frac{n(\mathrm{Var}(\mathsf{d}^{\max})+\mathbb{E}[\mathsf{d}^{\max}])}{\mathrm{Var}(\mathsf{d}^{\max})+\mathbb{E}[\mathsf{d}^{\max}](\mathbb{E}[\mathsf{d}^{\max}]+1)} \\
\underline{c}_u^{\mathrm{SS-C}} &\equiv \frac{n+1}{2+\mathbb{E}[\mathsf{d}^{\max}]} \\
\bar{c}_u^{\mathrm{SS-C}} &\equiv \frac{(n+1)(\mathrm{Var}(\mathsf{d}^{\max})+2\mathbb{E}[\mathsf{d}^{\max}])}{2(\mathrm{Var}(\mathsf{d}^{\max})+\mathbb{E}[\mathsf{d}^{\max}](2+\mathbb{E}[\mathsf{d}^{\max}]))}.
\end{aligned} \tag{6.20}$$

Under Scenario ii), where the target set $\mathcal{V}_0^* \subseteq \mathcal{V}_0$ *is independent of* G_0 *and has cardinality* $n_0^* = |\mathcal{V}_0^*|$*:*

$$\begin{aligned}
\underline{c}_u^{\mathrm{SS-R}} &\equiv \frac{n}{n+\bar{s}^{n_0^*}(n_0^*-n)} \\
\bar{c}_u^{\mathrm{SS-R}} &\equiv \frac{n(\bar{s}^{n_0^*}+n_0^*)}{n_0^*(n+\bar{s}^{n_0^*}(1+n_0^*-n))} \\
\underline{c}_u^{\mathrm{SS-C}} &\equiv \frac{n+1}{1+n+\bar{s}^{n_0^*}(n_0^*-n)} \\
\bar{c}_u^{\mathrm{SS-C}} &\equiv \frac{(n+1)(\bar{s}^{n_0^*}+1+n_0^*)}{(n_0^*+1)(1+n+\bar{s}^{n_0^*}(1+n_0^*-n))}.
\end{aligned} \tag{6.21}$$

Proof. Scenario *i*). The SS-R bounds are derived by applying Eq. (6.17) in Fact 3 to $n\mathbb{E}\left[\frac{1}{\mathsf{n}_0^{e,*}}\right]$, where $\mu = 1+\mathbb{E}[\mathsf{d}^{\max}]$, $\sigma^2 = \mathrm{Var}(\mathsf{d}^{\max})$, and $x_{\min} = 1$, with $\mathbb{E}[\mathsf{d}^{\max}]$ and $\mathrm{Var}(\mathsf{d}^{\max})$ in Eq. (6.13). The SS-C bounds are derived by applying Eq. (6.17) in Fact 3 to $(n+1)\mathbb{E}\left[\frac{1}{\mathsf{n}_0^{e,*}+1}\right]$, where $\mu = 2+\mathbb{E}[\mathsf{d}^{\max}]$, $\sigma^2 = \mathrm{Var}(\mathsf{d}^{\max})$, and $x_{\min} = 2$.

Scenario *ii*). The SS-R bounds are derived by applying Eq. (6.18) in Fact 3 to

$$n\mathbb{E}\left[(n_0^* + \mathrm{Bin}(n - n_0^*, 1 - \bar{s}^{n_0^*}))^{-1}\right] \tag{6.22}$$

i.e., with $a = n_0^*$, $m = n - n_0^*$, and $p = 1 - \bar{s}^{n_0^*}$. The SS-C bounds are derived by applying Eq. (6.18) in Fact 3 to $(n+1)\mathbb{E}\left[(n_0^* + 1 + \mathrm{Bin}(n - n_0^*, 1 - \bar{s}^{n_0^*}))^{-1}\right]$, i.e., with $a = n_0^* + 1$, $m = n - n_0^*$, and $p = 1 - \bar{s}^{n_0^*}$. □

6.3.2 SS-S

Throughout this subsection the assumption is retained that the initial graph is an ER random graph $\mathsf{G}_0 = (\mathcal{V}_0, \mathsf{E}_0)$ with parameters (n, s), and that the expectation is with respect to both the graph and sampling distributions. Recall that a star will hit the target set if its star center is in the extended neighborhood of the target. It follows that the (random) probability that star $t+1$ hits the target set is

$$\mathsf{p}_{t+1}^{\mathrm{SS-S}} \equiv \frac{\mathsf{n}_t^{e,*}}{\mathsf{n}_t}. \tag{6.23}$$

Define events $(\mathcal{C}_t, t \in \mathbb{Z}^+)$, with $\mathcal{C}_0$ trivial, and

$$\mathcal{C}_t \equiv \left\{\mathsf{v}_t \notin \mathsf{V}_{t-1}^{e,*})\right\}. \tag{6.24}$$

Here, $\mathcal{C}_t$ is the event that the sample t star misses the target set. Next, define events $(\bar{\mathcal{C}}_t, t \in \mathbb{Z}^+)$, with $\bar{\mathcal{C}}_0$ trivial, and

$$\bar{\mathcal{C}}_t \equiv \bigcap_{t' \in [t]} \mathcal{C}_{t'} = \left\{\mathsf{v}_{t'} \notin \mathsf{V}_{t'-1}^{e,*},\ \forall t' \in [t]\right\}. \tag{6.25}$$

Thus, $\bar{\mathcal{C}}_t$ is the event that the stars of the first t samples have each missed the target set. Observe that, conditioned on $\bar{\mathcal{C}}_t$, the target set in graph G_t is identical to the same set in the initial graph G_0, i.e., $\mathsf{V}_t^* | \bar{\mathcal{C}}_t = \mathcal{V}_0^*$, although the degrees of vertices in $\mathcal{V}_0^*$ may have decreased due to sampling.

The expected probability of hitting the target with the star drawn sample $t+1$, conditioned on missing the target set in the first t samples is:

$$\mathbb{E}[\mathsf{p}_{t+1}^{\mathrm{SS-S}} | \bar{\mathcal{C}}_t] = \mathbb{E}\left[\frac{\mathsf{n}_t^{e,*}}{\mathsf{n}_t}\middle| \bar{\mathcal{C}}_t\right]. \tag{6.26}$$

Leveraging Eq. (6.12), the error in approximating this expected conditional probability by its ratio

of expectations, i.e.,

$$\tilde{p}_{t+1}^{\mathrm{SS-S}} \equiv \frac{\mathbb{E}[\mathsf{n}_t^{e,*}|\bar{\mathcal{C}}_t]}{\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t]}, \tag{6.27}$$

is approximately

$$\epsilon_{n,t} \equiv \left| \frac{\mathrm{Var}(\mathsf{n}_t|\bar{\mathcal{C}}_t)\mathbb{E}[\mathsf{n}_t^{e,*}|\bar{\mathcal{C}}_t]}{\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t]^3} - \frac{\mathrm{Cov}(\mathsf{n}_t^{e,*}, \mathsf{n}_t|\bar{\mathcal{C}}_t)}{\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t]^2} \right|. \tag{6.28}$$

This approximate conditional hitting probability and its approximation error are expressed in terms of the parameters $(n, \mathbb{E}[\mathsf{n}_0^{e,*}], n_0^*, s, t)$ in the following theorem.

Theorem 13. *The approximate probability of hitting the target set in sample $t+1$ under SS-S, conditioned on missing the target set in the first t samples, is:*

$$\tilde{p}_{t+1}^{\mathrm{SS-S}} = \frac{(\mathbb{E}[\mathsf{n}_0^{e,*}] - n_0^*)\bar{s}^t + n_0^*}{n\bar{s}^t - \frac{\bar{s}}{s}(1-\bar{s}^t)}, \tag{6.29}$$

where $\mathbb{E}[\mathsf{n}_0^{e,*}]$ depends upon the search scenario.

Viewing $\tilde{p}_{t+1}^{\mathrm{SS-S}}$ as a continuous function of t, $\tilde{p}_{t+1}^{\mathrm{SS-S}}$ is convex increasing in t over $t \in [0, t_{\tilde{p}}^{(2)})$, where

$$t_{\tilde{p}}^{(2)} \equiv \frac{\log\left(n\frac{s}{\bar{s}} + 1\right)}{\log(1/\bar{s})}. \tag{6.30}$$

Moreover, $\tilde{p}_1^{\mathrm{SS-S}}$ (hitting in the first sample) equals the exact value $p_1^{\mathrm{SS-S}} \equiv n_0^{e,*}/n$, and $\tilde{p}_{t+1}^{\mathrm{SS-S}} = 1$ at

$$t_{\tilde{p}}^{(1)} \equiv \frac{\log\left(s(n - \mathbb{E}[\mathsf{n}_0^{e,*}] + n_0^*) + \bar{s}\right) - \log\left(\bar{s} + n_0^* s\right)}{\log(1/\bar{s})} < t_{\tilde{p}}^{(2)}. \tag{6.31}$$

The error in this approximation, $\epsilon_{n,t}$, has upper bound

$$\epsilon_{n,t} \leq \left(n(1-\bar{s}^t) + \frac{\bar{s}}{s(1+\bar{s})}\left(1 + \bar{s}(1-\bar{s}^t)\right)\right)\bar{s}^t \times \frac{(\mathbb{E}[\mathsf{n}_0^{e,*}] - n_0^*)\bar{s}^t + n_0^*}{\left(n\bar{s}^t - \frac{\bar{s}}{s}(1-\bar{s}^t)\right)^3}. \tag{6.32}$$

The approximation error is asymptotically negligible in n:

$$\lim_{n\uparrow\infty} \epsilon_{n,t} = 0, \ \forall t \in \mathbb{N}, s \in (0,1). \tag{6.33}$$

and in fact $\epsilon_{n,t} = O(\mathbb{E}[\mathsf{n}_0^{e,*}]/n^2)$.

Proof. Eq. (6.29) *through Eq.* (6.31). Apply Lem. 20 in Sec. 6.11 to each of the numerator and denominator in Eq. (6.27). Specifically, $\mathbb{E}[\mathsf{n}_t^{e,*}|\bar{\mathcal{C}}_t]$ follows from the expectation expression for case i) with watch set $\Gamma^e(\mathcal{V}^*)$, (disjoint) draw set $\mathcal{V}_0 \setminus \Gamma^e(\mathcal{V}^*)$, conditioned on $\bar{\mathcal{E}}$ and hence $\mathcal{V}^*$ being

unsampled, while $\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t]$ follows from the expectation expression for case *ii*) with watch set $\mathcal{V}_0$ and (subset) draw set $\mathcal{V}_0 \setminus \Gamma^e(\mathcal{V}^*)$:

$$\mathbb{E}[\mathsf{n}_t^{e,*}|\bar{\mathcal{C}}_t] = (\mathbb{E}[\mathsf{n}_0^{e,*}] - n_0^*)\bar{s}^t + n_0^* \tag{6.34}$$

$$\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t] = n\bar{s}^t - \frac{\bar{s}}{s}(1-\bar{s}^t). \tag{6.35}$$

The ratio of Eq. (6.34) and Eq. (6.35) gives Eq. (6.29). The first two derivatives of $\tilde{p}_t^{\mathrm{SS-S}}$ with respect to t are:

$$\tilde{p}_{t+1}^{\mathrm{SS-S}'} = \frac{s\bar{s}^t(n_0^* ns + \mathbb{E}[\mathsf{n}_0^{e,*}]\bar{s})\log(\frac{1}{\bar{s}})}{(\bar{s}-\bar{s}^t(ns+\bar{s}))^2} > 0$$

$$\tilde{p}_{t+1}^{\mathrm{SS-S}''} = \frac{ss^{t+1}\mathbb{E}[\mathsf{n}_0^{e,*}](\bar{s}+\bar{s}^t(ns+\bar{s}))\,(\log\bar{s})^2}{(\bar{s}^t(ns+\bar{s})-\bar{s})^3} > 0.$$

This establishes that $\tilde{p}_t^{\mathrm{SS-S}}$ is convex increasing. The value $t_{\tilde{p}}^{(2)}$ is found by equating the denominator in $\tilde{p}_t^{\mathrm{SS-S}}$ to zero and solving for t, and $t_{\tilde{p}}^{(1)}$ is found by solving $\tilde{p}_t^{\mathrm{SS-S}} = 1$ for t.

Proof of Eq. (6.32). Apply the variance expression in case *ii*) of Lem. 20 in Sec. 6.11 with watch set $\mathcal{V}_0$ and (subset) draw set $\mathcal{V}_0 \setminus \Gamma^e(\mathcal{V}^*)$ to obtain:

$$\mathrm{Var}(\mathsf{n}_t|\bar{\mathcal{C}}_t) = \left(n(1-\bar{s}^t) + \frac{\bar{s}}{s(1+\bar{s})}\left(1+\bar{s}(1-\bar{s}^t)\right)\right)\bar{s}^t. \tag{6.36}$$

Next, observe $\mathrm{Cov}(\mathsf{n}_t^{e,*}, \mathsf{n}_t|\bar{\mathcal{C}}_t)$ in Eq. (6.28) is nonnegative. To see this, consider a general scenario where each member of a population is given a property value, and two subsets of the population are defined as holding those members with property values in a given target set, with the second target set a subset of the first. Let the property values be random, and consider the two random variables denoting the cardinalities of the two random population subsets. These random variables are by construction positively correlated. This general scenario applies here to the population V_t with the first subset equal to V_t and the second subset equal to $\Gamma^e(\mathsf{V}_t^*)$.

By the above argument, Eq. (6.28) may be upper bounded as:

$$\epsilon_{n,t} \leq \frac{\mathrm{Var}(\mathsf{n}_t|\bar{\mathcal{C}}_t)\mathbb{E}[\mathsf{n}_t^{e,*}|\bar{\mathcal{C}}_t]}{\mathbb{E}[\mathsf{n}_t|\bar{\mathcal{C}}_t]^3}. \tag{6.37}$$

Substitution of Eq. (6.34), Eq. (6.35), and Eq. (6.36) into Eq. (6.37) gives Eq. (6.32).

Proof of Eq. (6.33). This follows immediately by observing the numerator is $O(n\mathbb{E}[\mathsf{n}_0^{e,*}])$ while the denominator is $O(n^3)$. □

Corollary 4. *Let* $(\mu_{\mathsf{n},t}, t \in \mathbb{N})$*, with* $\mu_{\mathsf{n},t} \equiv \mathbb{E}[\mathsf{n}_t]$*, and* $(\sigma^2_{\mathsf{n},t}, t \in \mathbb{N})$*, with* $\sigma^2_{\mathsf{n},t} \equiv \mathrm{Var}(\mathsf{n}_t)$*, denote the means and variances of* $(\mathsf{n}_t, t \in \mathbb{N})$*. Then,* $\mu_{\mathsf{n},t}$ *is given by Eq.* (6.35) *and* $\sigma^2_{\mathsf{n},t}$ *is given by Eq.* (6.36)*.*

The function $\mu_{\mathsf{n},t}$ *is strictly decreasing in* t*, and intersects the* t*-axis at*

$$t^*_\mu = \frac{1}{\log \bar{s}} \log \frac{\bar{s}}{ns + \bar{s}}. \tag{6.38}$$

The function $\sigma^2_{\mathsf{n},t}$ *is concave increasing in* t *over* $[0, t^{(1)}_{\sigma^2}]$*, concave decreasing over* $(t^{(1)}_{\sigma^2}, t^{(2)}_{\sigma^2}]$*, convex decreasing over* $(t^{(2)}_{\sigma^2}, \infty)$*, and intersects the* t*-axis at* $t^{(3)}_{\sigma^2}$*, where*

$$0 < t^{(1)}_{\sigma^2} < t^{(2)}_{\sigma^2} < t^{(3)}_{\sigma^2} \tag{6.39}$$

and

$$\begin{aligned}
t^{(1)}_{\sigma^2} &= \frac{1}{\log \bar{s}} \log \frac{(1+\bar{s})(ns+\bar{s})}{2((n-1)(1-\bar{s}^2)+1)} \\
t^{(2)}_{\sigma^2} &= \frac{1}{\log \bar{s}} \log \frac{(1+\bar{s})(ns+\bar{s})}{4((n-1)(1-\bar{s}^2)+1)} \\
t^{(3)}_{\sigma^2} &= \frac{1}{\log \bar{s}} \log \frac{\bar{s}}{(n-1)(1-\bar{s}^2)+1}.
\end{aligned} \tag{6.40}$$

Proof. The first two derivatives of $\mu_{\mathsf{n},t}$ with respect to t are:

$$\begin{aligned}
\mu'_{\mathsf{n},t} &= \frac{ns+\bar{s}}{s}(\log \bar{s})\bar{s}^t < 0 \\
\mu''_{\mathsf{n},t} &= \frac{ns+\bar{s}}{s}(\log \bar{s})^2\bar{s}^t > 0.
\end{aligned} \tag{6.41}$$

The first two derivatives of $\sigma^2_{\mathsf{n},t}$ with respect to t are:

$$\begin{aligned}
\left(\sigma^2_{\mathsf{n},t}\right)' &= \frac{\bar{s}^t}{1-\bar{s}^2}\left[(1+\bar{s})(ns+\bar{s}- \right. \\
&\qquad \left. 2((n-1)(1-\bar{s}^2)+1)\bar{s}^t\right]\log \bar{s} \\
\left(\sigma^2_{\mathsf{n},t}\right)'' &= \frac{\bar{s}^t}{1-\bar{s}^2}\left[(1+\bar{s})(ns+\bar{s}- \right. \\
&\qquad \left. 4((n-1)(1-\bar{s}^2)+1)\bar{s}^t\right](\log \bar{s})^2.
\end{aligned} \tag{6.42}$$

Equating these with zero yields the given expressions. □

The theorem below approximates both the *unconditional* probability of first hitting the target set in sample $t+1$, and the expected unit cost under SS-S.

Theorem 14. *The approximate* unconditional *probability of first hitting the target set in sample t under SS-S is*

$$\tilde{q}_t^{\mathrm{SS-S}} \equiv \tilde{p}_t^{\mathrm{SS-S}} \prod_{u\in[t-1]} (1-\tilde{p}_u^{\mathrm{SS-S}}), \tag{6.43}$$

for $t \in [\lfloor t_{\tilde{p}}^{(1)} \rfloor]$, with $t_{\tilde{p}}^{(1)}$ defined in Eq. (6.31)*. Letting $\tilde{\mathsf{c}}_u^{\mathrm{SS-S}}$ denote the approximate random unit cost under SS-S sampling, i.e., with the RV $\tilde{\mathsf{c}}_u^{\mathrm{SS-S}}$ drawn according to the distribution $\tilde{q}^{\mathrm{SS-S}} \equiv (\tilde{q}_t^{\mathrm{SS-S}}, t \in [t_{\tilde{p}}^{(1)}])$, the associated approximate* expected *unit cost under SS-S sampling is*

$$\tilde{c}_u^{\mathrm{SS-S}} \equiv \sum_{t\in[t_{\tilde{p}}^{(1)}]} \prod_{u\in[t]} (1-\tilde{p}_u^{\mathrm{SS-S}}). \tag{6.44}$$

The proof is found in Sec. 6.11.

Numerical results for p_t and q_t under SS-R, SS-C, and SS-S sampling on ER graphs are shown in Fig. 6.3 for Scenario *i*) and Fig. 6.4 for Scenario *ii*). Under Scenario *i*) there is a clear offset in the estimates of the conditioned probability, p_t, for all variants of star sampling as a result of the underestimate of $\mathbb{E}[\mathsf{n}_0^{e,*}]$, see Fig. 6.2. This offset leads to the error in the estimate of the unconditioned probability q_t observed in Scenario *i*), the bottom plot of Fig. 6.3. Under Scenario *ii*) these estimates are fairly accurate. Additionally, Fig. 6.4 shows that the ordering of p_t and q_t under the three sampling variants is reflected in the ordering of the estimates of p_t and q_t.

Numerical results for the expected unit cost to find $v \in \mathcal{V}^*$ on ER graphs under SS-R, SS-C, and SS-S and the estimates $[\underline{c}_u^{\mathrm{SS-R}}, \bar{c}_u^{\mathrm{SS-R}}]$, $[\underline{c}_u^{\mathrm{SS-C}}, \bar{c}_u^{\mathrm{SS-C}}]$, and $\tilde{c}_u^{\mathrm{SS-S}}$ are given in Fig. 6.5 for Scenario *i*) and Fig. 6.6 for Scenario *ii*). Under both Scenario *i*) and *ii*) the estimated unit costs are fairly close to the empirical unit costs which in turn are nearly identical for $n = 1000$.

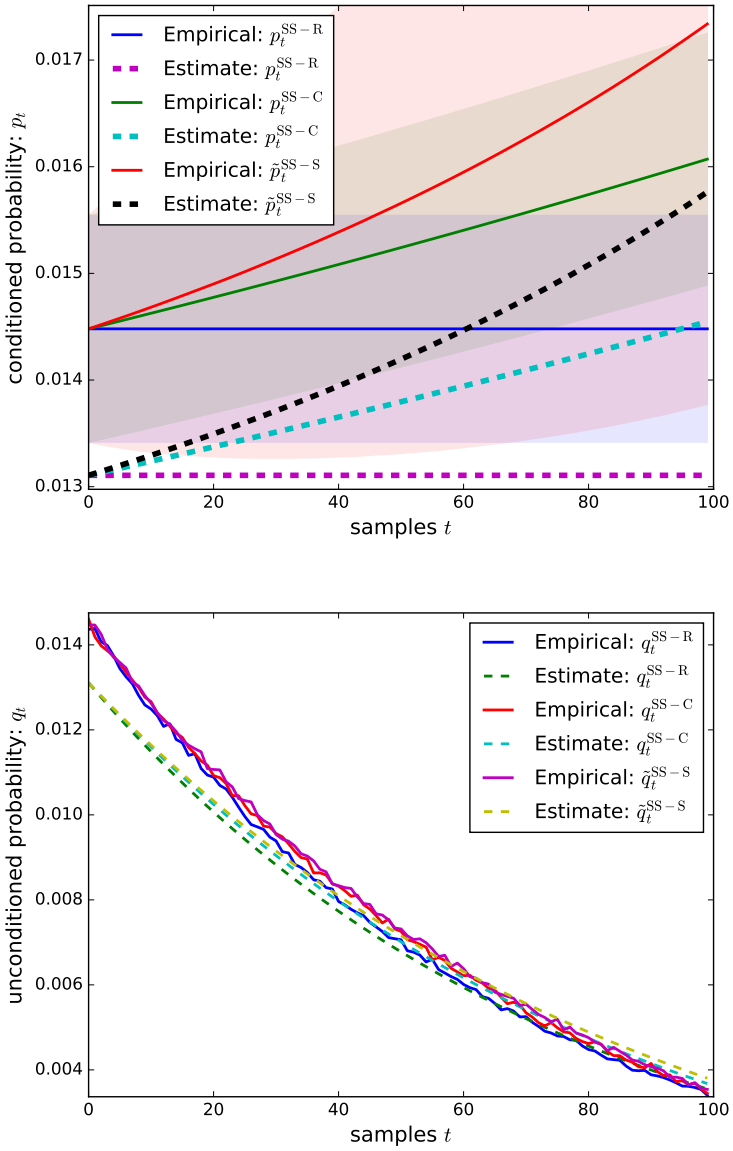

Figure 6.3: Conditional p_t (top) and unconditional q_t (bottom) hit probability for Scenario i): $n = 1000$, $s = 0.005$, 20k trials, 100 graphs, shaded regions represent empirical standard deviation.

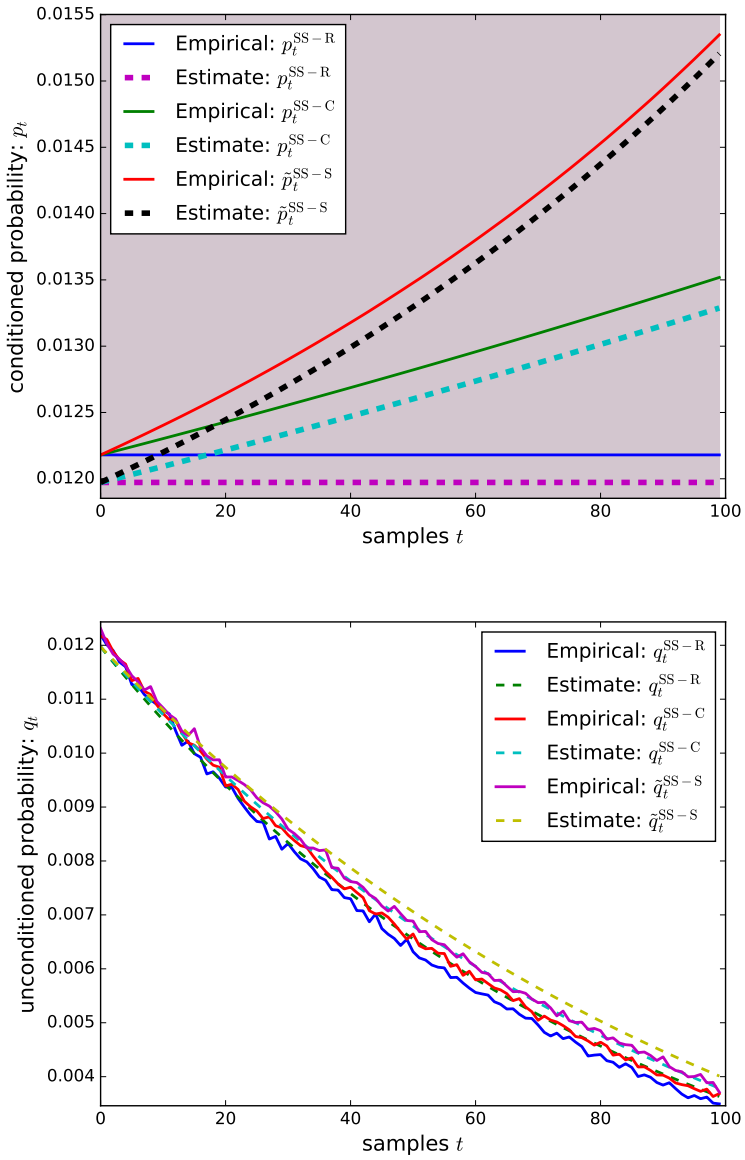

Figure 6.4: Conditional p_t (top) and unconditional q_t (bottom) hit probability for Scenario *ii*): $n = 1000$, $s = 0.005$, 20k trials, 100 graphs, $n_0^* = 2$, shaded regions represent empirical standard deviation.

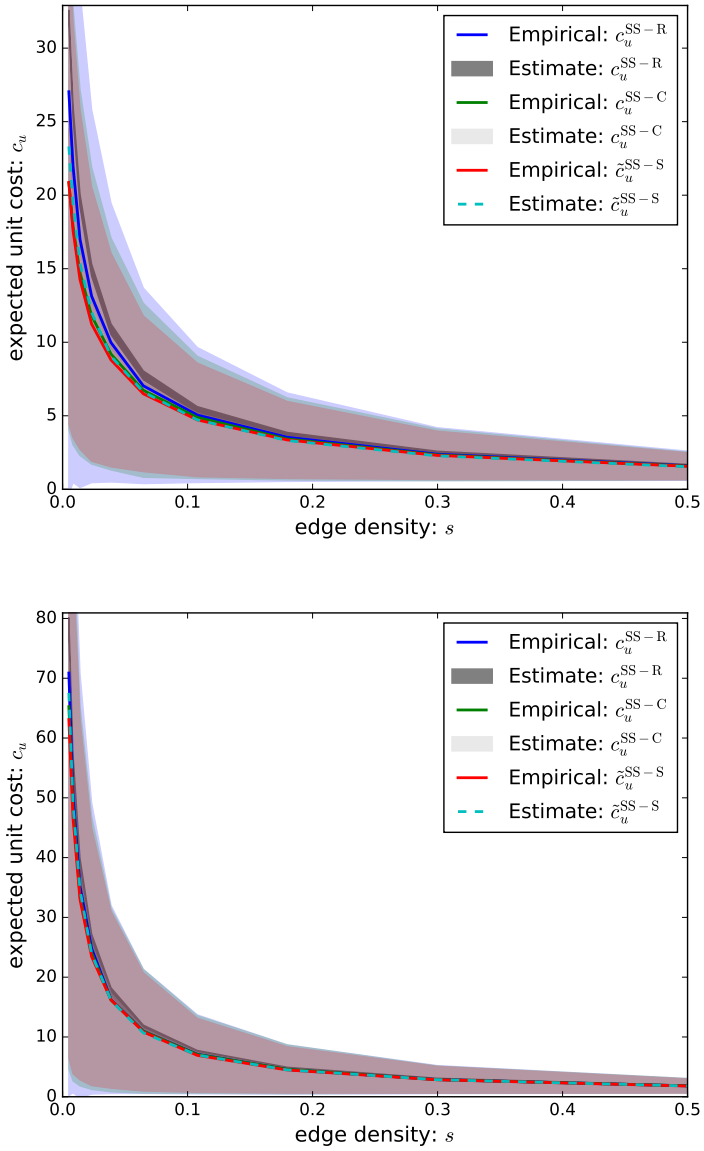

Figure 6.5: Expected unit cost for Scenario *i*) (top) and *ii*) (bottom): $n = 1000$, $s = 0.005$, 100 trials, 100 graphs, shaded regions represent empirical standard deviation.

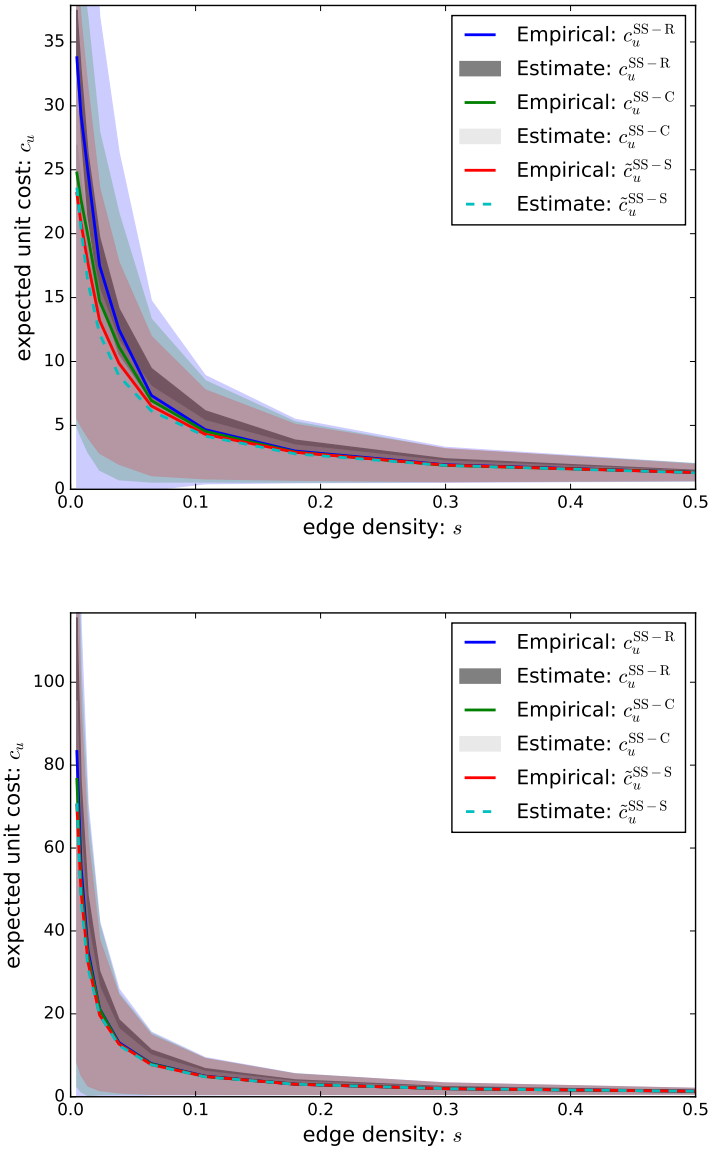

Figure 6.6: Expected unit cost for Scenario *ii*) with initial ER random graphs of order $n = 100$ (top) and $n = 1000$ (bottom): $s = 0.005$, 100 trials, 100 graphs, $n_0^* = 2$, shaded regions represent empirical standard deviation.

Given the observation in Fig. 6.5 and Fig. 6.6 of similar expected unit costs of SS-R, SS-C, and SS-S, it is desirable to compare the approximate expected unit cost under SS-S (in Eq. (6.44)) with (the bounds on) the expected unit cost under SS-R and SS-C (in Prop. 15). However, the dependence of the SS-S cost upon the underlying parameters $(n, s, \mathbb{E}[\mathsf{n}_0^{e,*}])$ is too complex for such a comparison to be insightful. Consequently, this chapter instead focuses on showing in Prop. 16 that *i*) the approximate *conditional* probability of hitting the target set for the first time in sample

t under SS-S, $\tilde{p}_t^{\mathrm{SS-S}}$, is close to that under SS-R, $p_t^{\mathrm{SS-R}}$; *ii*) $p_t^{\mathrm{SS-R}}$ is close to $p_t^{\mathrm{SS-C}}$; and (therefore) *iii*) $\tilde{p}_t^{\mathrm{SS-S}}$ is close to $p_t^{\mathrm{SS-C}}$. This is shown by considering the ER edge probability s to be $O(1/n)$, e.g., $s(n) = c/n$, for some $c > 0$.

Proposition 16. *Letting $p_t^{\mathrm{SS-R}}$, $p_t^{\mathrm{SS-C}}$ be the probability of sampling the target set for the first time on sample t in SS-R and SS-C, and $\tilde{p}_t^{\mathrm{SS-S}}$ be the approximate probability of sampling the target set for the first time on sample t in SS-S. For $s(n) = \frac{c}{n}$ (with $c > 0$) and finite $t \geq 1$, as $n \to \infty$,*

$$\frac{\tilde{p}_t^{\mathrm{SS-S}}}{p_t^{\mathrm{SS-R}}} \to 1, \ \frac{p_t^{\mathrm{SS-R}}}{p_t^{\mathrm{SS-C}}} \to 1, \ \frac{\tilde{p}_t^{\mathrm{SS-S}}}{p_t^{\mathrm{SS-C}}} \to 1, \tag{6.45}$$

under both Scenario i) and Scenario ii).

The proof of Prop. 16 is given in Sec. 6.11.

6.4 Relative unit cost of SS-R, SS-C, and SS-S

The purpose of this section is to provide examples demonstrating that, in contrast with SS-C and SS-R (c.f. Rem. 3), there is no guaranteed ordering of the expected unit costs of *i*) SS-C and SS-S, or *ii*) SS-R and SS-S, for all graphs and all choices of target set.

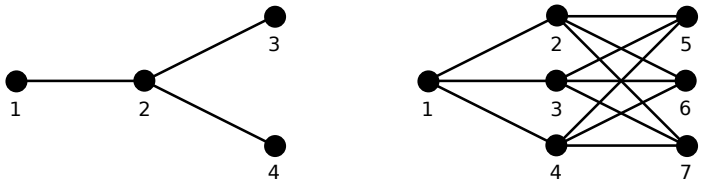

Figure 6.7: Left: graph $G^{(1)}$ with target set $\mathcal{V}^* \equiv \{1\}$, for which SS-C outperforms SS-S. Right: $G^{(2)}$ with target set $\mathcal{V}^* \equiv \{1\}$, for which SS-R outperforms SS-S.

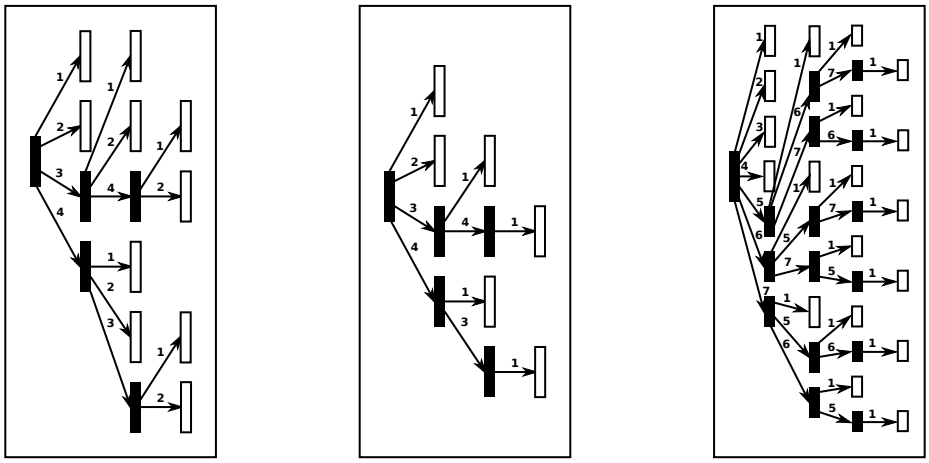

Figure 6.8: Left: outcome tree for SS-C on $G^{(1)}$. Center: outcome tree for SS-S on $G^{(1)}$. Right: outcome tree for SS-S on $G^{(2)}$. White blocks are terminating states.

6.4.1 SS-C may outperform SS-S

The expected performance under SS-C and SS-S on a given graph may be analyzed by making outcome trees, as illustrated in Fig. 6.8 for the graphs in Fig. 6.7. Each level of the tree corresponds to a time instant $t \in \mathbb{Z}_+$, with the root vertex (corresponding to $t = 0$) representing the initial graph $G_0 = (\mathcal{V}_0, \mathcal{E}_0)$. Each vertex in the tree at level t corresponds to a particular graph possible at time t. Each labeled edge in the tree, connecting a graph at time t with a graph at time $t + 1$, corresponds to a choice of the star center at sample $t \in \mathbb{N}$. For each vertex v in the tree, corresponding to, say, a graph $G(v) = (\mathcal{V}(v), \mathcal{E}(v))$, there is a collection of edges in the tree, emanating from v, one edge for each vertex in $G(v)$, corresponding to the possible star centers that may be chosen from $G(v)$. Each of these edges has a probability of $1/|\mathcal{V}(v)|$, as each vertex in $\mathcal{V}(v)$ is equally likely to be selected.

Leaf vertices in the tree are terminating states, representing the fact that the target set $\mathcal{V}^*$ has been hit for the first time. Let $\bar{\mathcal{L}}$ and $\breve{\mathcal{L}}$ denote the leaves in the outcome trees for a given graph under SS-C and SS-S, respectively. Each leaf has a unique path to the root vertex, and the probability of the leaf is the product of the probabilities assigned to the edges comprising that path. Define $\bar{\mathbf{p}}_{\mathcal{L}} \equiv (\bar{P}_L, L \in \bar{\mathcal{L}})$ and $\breve{\mathbf{p}}_{\mathcal{L}} \equiv (\breve{P}_L, L \in \breve{\mathcal{L}})$ as the probability distributions on the leaves of the outcome trees under SS-C and SS-S, respectively. Finally, observe each leaf L has a depth, denoted $\bar{N}_L, \breve{N}_L \in \mathbb{N}$, i.e., a length of the path from the root, and this corresponds to the number of samples until the target set was hit. It follows that

$$c_u^{\mathrm{SS-C}} = \sum_{L \in \bar{\mathcal{L}}} \bar{N}_L \bar{P}_L, \qquad c_u^{\mathrm{SS-S}} = \sum_{L \in \breve{\mathcal{L}}} \breve{N}_L \breve{P}_L. \tag{6.46}$$

Fact 4 (SS-C may outperform SS-S)**.** *There exist graphs and target sets for which the expected unit cost of SS-C outperforms that of SS-R, i.e.,* $c_u^{\mathrm{SS-C}} < c_u^{\mathrm{SS-S}}$.

Proof. Fix $G^{(1)}$ in Fig. 6.7 and fix the target set $\mathcal{V}^* = \{1\}$. The outcome tree shown in the left of Fig. 6.8, corresponding to running SS-C on $G^{(1)}$, has

$\bar{M}_L$	$\bar{N}_L$	$\bar{P}_L$	$\bar{M}_L\bar{N}_L\bar{P}_L$
2	1	1/4	1/2
4	2	1/12	2/3
4	3	1/24	1/2

(6.47)

where $\bar{M}_L$ is the number of leaf vertices of type $(\bar{N}_L, \bar{P}_L)$. Adding up the right column gives $c_u^{\mathrm{SS-C}} = 5/3$. The outcome tree shown in the middle of Fig. 6.8, corresponding to running SS-S on $G^{(1)}$, has

$\breve{M}_L$	$\breve{N}_L$	$\breve{P}_L$	$\breve{M}_L\breve{N}_L\breve{P}_L$
2	1	1/4	1/2
2	2	1/8	1/2
2	3	1/8	3/4

(6.48)

Summation gives $c_u^{\mathrm{SS-S}} = \frac{7}{4}$. Thus $\frac{5}{3} = c_u^{\mathrm{SS-C}} < c_u^{\mathrm{SS-S}} = \frac{7}{4}$. □

6.4.2 SS-R may outperform SS-S

Fact 5 (SS-R may outperform SS-S)**.** *There exist graphs for which star sampling with replacement outperforms star sampling without replacement of star, i.e.,* $c_u^{\mathrm{SS-R}} < c_u^{\mathrm{SS-S}}$.

Proof. Fix $G^{(2)}$ in Fig. 6.7 and fix the target set $\mathcal{V}^* = \{1\}$. Note $n_0^{e,*} = 4$ and $n = 7$, and thus the performance under SS-R is, by Fact 1, $c_u^{\mathrm{SS-R}} = 7/4$. The outcome tree shown in the right of Fig. 6.8, corresponding to running SS-S on $G^{(2)}$, has

$\breve{M}_L$	$\breve{N}_L$	$\breve{P}_L$	$\breve{M}_L\breve{N}_L\breve{P}_L$
4	1	1/7	4/7
3	2	1/21	5/21
6	3	1/42	3/7
6	4	1/42	4/7

(6.49)

Summation gives $c_u^{\mathrm{SS-S}} = \frac{38}{21}$. Thus $\frac{7}{4} = c_u^{\mathrm{SS-R}} < c_u^{\mathrm{SS-S}} = \frac{38}{21}$. □

Remark 4. *The performance of SS-S is worse than SS-C and SS-R, respectively, in the two previous examples. This may be counter-intuitive, given that SS-S removes more vertices outside the target set than the other two. However, as these examples show, vertices outside the target set include the* neighbors *of the target set, and removing them may hurt the expected performance, as the target set is harder to "hit" with a randomly selected star when it has fewer neighbors.*

6.5 Linear Cost Model

The expected linear cost (Def. 24) of SS-R on an arbitrary graph is in Fact 6. The approximate expected linear costs of SS-R, SS-C, and SS-S on an ER random graph are in Sec. 6.5.2.

6.5.1 Arbitrary graph

Let $G = (\mathcal{V}, \mathcal{E})$ be an arbitrary graph with order $|\mathcal{V}| = n$, and let $\mathcal{V}^* \subseteq \mathcal{V}$ be an arbitrary target set. Fact 6 below is the linear cost analog of the unit cost result for SS-R in Fact 1.

This section first develops some necessary notation, extending that introduced in Sec. 6.2.1. Let $\mathcal{V}_G^{e,*} = \Gamma^e(\mathcal{V}^*)$ denote the extended neighborhood of $\mathcal{V}^*$ (with order $n_G^{e,*} \equiv |\mathcal{V}_G^{e,*}|$), and let $\bar{\mathcal{V}}_G^{e,*} \equiv \mathcal{V} \setminus \mathcal{V}_G^{e,*}$ denote its complement. Define $\mathcal{V}_G^{e,*}(k) \equiv \{v \in \mathcal{V}_G^{e,*} : d_v = k\}$ (for $k \in \mathcal{V}_G^{e,*}$) and $\bar{\mathcal{V}}_G^{e,*}(k) \equiv \{v \in \bar{\mathcal{V}}_G^{e,*} : d_v = k\}$ (for $k \in \bar{\mathcal{V}}_G^{e,*}$) as the subsets of vertices in $\mathcal{V}_G^{e,*}, \bar{\mathcal{V}}_G^{e,*}$, respectively, of degree k. The conditional degree distributions in $\mathcal{V}_G^{e,*}, \bar{\mathcal{V}}_G^{e,*}$ are denoted $w_G^{e,*}, \bar{w}_G^{e,*}$, respectively, with components

$$w_G^{e,*}(k) \equiv \frac{|\mathcal{V}_G^{e,*}(k)|}{|\mathcal{V}_G^{e,*}|},\ \bar{w}_G^{e,*}(k) \equiv \frac{|\bar{\mathcal{V}}_G^{e,*}(k)|}{|\bar{\mathcal{V}}_G^{e,*}|}. \tag{6.50}$$

Finally, the average degrees in $\mathcal{V}_G^{e,*}, \bar{\mathcal{V}}_G^{e,*}$ with degree sets $\mathcal{D}^{e,*}, \bar{\mathcal{D}}^{e,*}$ respectively, are

$$d_G^{e,*} = \sum_{k \in \mathcal{D}_G^{e,*}} k w_G^{e,*}(k),\ \bar{d}_G^{e,*} = \sum_{k \in \bar{\mathcal{D}}_G^{e,*}} k \bar{w}_G^{e,*}(k). \tag{6.51}$$

Fact 6 (Linear cost of SS-R)**.** *Under SS-R, for any graph G and any target set $\mathcal{V}^*$, the expected linear cost c_l in Eq.* (6.9) *is*

$$c_l^{\mathrm{SS-R}}(G, \mathcal{V}^*) = \bar{d}_G^{e,*} \left(\frac{n}{n_G^{e,*}} - 1 \right) + d_G^{e,*}. \tag{6.52}$$

Proof. Let $p_t^{\mathrm{SS-R}} \equiv n_G^{e,*}/n$ denote the probability of success under SS-R, with $n_G^{e,*} \equiv |\mathcal{V}_G^{e,*}|$. Recall from Def. 23 that $\mathsf{c}_u = \mathsf{c}_u^{\mathrm{SS-R}}(G, \mathcal{V}^*)$, the random *unit* cost, denotes the random number of samples until success, and recall from Fact 1 that this quantity has distribution geo $\left(p_t^{\mathrm{SS-R}}\right)$. The random *linear* cost under SS-R, denoted $\mathsf{c}_l^{\mathrm{SS-R}}(G, \mathcal{V}^*)$, equals $\mathsf{c}_l = \mathsf{x}_1 + \cdots + \mathsf{x}_{\mathsf{c}_u}$, where x_t is the random

extended degree of the star center in sample t. The expected linear cost is

$$\begin{aligned} c_l^{\mathrm{SS-R}}(G, \mathcal{V}^*) &= \mathbb{E}[\mathsf{c}_l^{\mathrm{SS-R}}(G, \mathcal{V}^*)] \\ &= \mathbb{E}[(\mathsf{x}_1 + \cdots + \mathsf{x}_{\mathsf{t}-1}) + \mathsf{x}_{\mathsf{c}_u}] \\ &= \mathbb{E}[\mathbb{E}[(\mathsf{x}_1 + \cdots + \mathsf{x}_{\mathsf{c}_u - 1}) + \mathsf{x}_{\mathsf{c}_u} | \mathsf{c}_u]] \\ &= \mathbb{E}[(\mathsf{c}_u - 1)\mathbb{E}[\mathsf{x}_1 | \mathsf{c}_u]] + \mathbb{E}[\mathbb{E}[\mathsf{x}_{\mathsf{c}_u} | \mathsf{c}_u]]. \end{aligned}$$

The first term represents the expected linear cost up until but not including the cost of the final sample, while the second term is the expected cost of the final sample. The unsuccessful samples are identically distributed, due to replacement. The expected cost of an unsuccessful sample is $\bar{d}_G^{e,*}$ while the expected cost of a successful sample is $d_G^{e,*}$. The distribution of the cost of a sample is conditionally independent of the number of samples, conditioned on whether or not the sample is successful or not. As such, $\mathbb{E}[\mathsf{x}_1 | \mathsf{c}_u] = \bar{d}_G^{e,*}$, $\mathbb{E}[\mathsf{x}_{\mathsf{c}_u} | \mathsf{c}_u] = d_G^{e,*}$, and $\mathbb{E}[\mathsf{c}_u] = n/n_G^{e,*}$, yielding Eq. (6.52). □

In spite of the success in deriving the previous result for SS-R, it appears that a simple exact expression for the expected linear cost under SS-C on an arbitrary graph is not available. In contrast to the *unit* cost case, for which the SS-C expected unit cost is given by Fact 2, the linear cost case appears to be much more difficult to analyze, as it requires a representation of the evolution under SS-C of the (effectively arbitrary) degree distribution w_G of the initial graph G. This evolution does not appear to be sufficiently tractable to yield "closed-form" results like those in Sec. 6.3.1. Given this difficulty, the next section turns to approximations for the case of an ER random graph.

6.5.2 ER random graph

Now consider the case where the initial graph is an ER random graph $\mathsf{G}_0 = (\mathcal{V}_0, \mathsf{E}_0)$ with parameters (n, s), and where the expectation is with respect to the graph and sampling distributions. The common starting point for all three sampling paradigms is to express the expected linear cost in terms of the conditional expectations of the extended degrees in each sample, conditioned on the unit cost, and to make the *approximation* that the per-sample extended degree is *independent* of the

unit cost:

$$\begin{aligned} c_l \equiv \mathbb{E}[\mathsf{c}_l] &= \mathbb{E}\left[\sum_{t\in[c_{\max}]} \mathsf{d}_t^e \mathbf{1}(t \le \mathsf{c}_u)\right] \\ &= \sum_{t\in[c_{\max}]} \mathbb{E}[\mathsf{d}_t^e \mathbf{1}(t \le \mathsf{c}_u)] \\ &\approx \sum_{t\in[c_{\max}]} \mathbb{E}[\mathsf{d}_t^e]\mathbb{P}(t \le \mathsf{c}_u). \end{aligned} \tag{6.53}$$

Here, d_t^e is the extended degree of sample t, c_u is the unit cost, i.e., the random number of samples required for the star center to hit the extended target set, and $c_{\max}$ is an upper bound, possibly infinite, on c_u. As shown in Fact 6 and its proof, the extended degree is *not* independent of the number of samples, as the target set in G_t will have a distinct degree distribution from its complement, conditioned on the previous $t-1$ samples missing the target. Nonetheless, the approximation is useful in that it facilitates analysis and the numerical investigations in this chapter support the claim that the approximation is accurate over a wide array of parameter values of practical interest. In what follows below the above approximation is adapted to give the approximate expected linear cost *conditioned* on the random extended target set, $\mathsf{n}_0^{e,*}$, i.e.,

$$\mathbb{E}[\tilde{\mathsf{c}}_l|\mathsf{n}_0^{e,*}] \approx \tilde{\mathbb{E}}[\tilde{\mathsf{c}}_l|\mathsf{n}_0^{e,*}] \equiv \sum_{t\in[c_{\max}(\mathsf{n}_0^{e,*})]} \mathbb{E}[\mathsf{d}_t^e]\mathbb{P}(t \le \mathsf{c}_u|\mathsf{n}_0^{e,*}). \tag{6.54}$$

The additional approximation in Eq. (6.54), beyond those in Eq. (6.53), is that d_t^e is independent of $\mathsf{n}_0^{e,*}$, i.e., $\mathbb{E}[\mathsf{d}_t^e|\mathsf{n}_0^{e,*}] = \mathbb{E}[\mathsf{d}_t^e]$.

SS-R

In the case of SS-R, the approximation in Eq. (6.54) leads to a particularly simple expression on account of the fact that (d_t^e) are IID. This implies $\mathbb{E}[\mathsf{d}_t^e]$ does not depend upon t, and as such Wald's identity can be leveraged for the expected value of the sum of a random number of IID random variables:

$$\begin{aligned} \tilde{\mathbb{E}}[\tilde{\mathsf{c}}_l^{\mathrm{SS-R}}|\mathsf{n}_0^{e,*}] &= \mathbb{E}[\mathsf{d}_1^e] \sum_{t\in[c_{\max}^{\mathrm{SS-R}}(\mathsf{n}_0^{e,*})]} \mathbb{P}(t \le \mathsf{c}_u|\mathsf{n}_0^{e,*}) \\ &= \mathbb{E}[\mathsf{d}_1^e]\mathbb{E}[\mathsf{c}_u|\mathsf{n}_0^{e,*}] \\ &= (1+(n-1)s)\frac{n}{\mathsf{n}_0^{e,*}}. \end{aligned} \tag{6.55}$$

Here, $\mathbb{E}[\mathsf{d}_1^e] = 1 + (n-1)s$ is the expected extended degree of a random vertex in an ER random graph, and $\mathbb{E}[\mathsf{c}_u|\mathsf{n}_0^{e,*}] = n/\mathsf{n}_0^{e,*}$ is the conditional expected unit cost of SS-R. Taking the expectation of the above with respect to $\mathsf{n}_0^{e,*}$ establishes the following proposition.

Proposition 17 (Linear cost of SS-R for ER graph.)**.** *Fix the initial graph as an ER random graph* G_0 *with parameters* (n, s)*, and the extended target set cardinality* $\mathsf{n}_0^{e,*}$*. The approximate expected linear cost under SS-R,* $\tilde{c}_l^{\mathrm{SS-R}}$*, is*

$$\tilde{c}_l^{\mathrm{SS-R}} \equiv \mathbb{E}[\mathbb{E}[\tilde{\mathsf{c}}_l^{\mathrm{SS-R}}|\mathsf{n}_0^{e,*}]] = (1 + (n-1)s)c_u^{\mathrm{SS-R}}, \tag{6.56}$$

and the SS-R expected unit cost, $c_u^{\mathrm{SS-R}}$*, has bounds in Prop. 15.*

SS-C

Leveraging the approximation in Eq. (6.54) for SS-C requires the expected extended degree of the random star center selected in sample t and the unit cost distribution.

Proposition 18 (Linear cost of SS-C for ER graph.)**.** *The approximate expected linear cost under SS-C for an ER graph, conditioned on the random extended target set order* $\mathsf{n}_0^{e,*}$*, is*

$$\mathbb{E}[\tilde{\mathsf{c}}_l^{\mathrm{SS-C}}|\mathsf{n}_0^{e,*}] = \sum_{t=1}^{n-\mathsf{n}_0^{e,*}+1} (1 + (n-t)s) \prod_{u=1}^{t-1} \left(1 - \frac{\mathsf{n}_0^{e,*}}{n-u+1}\right). \tag{6.57}$$

Proof. Consider an urn with n balls of which k are marked. Draw m balls uniformly at random and let x be the random number of marked balls drawn, with support $\mathcal{S} \equiv \{\max\{0, m-(n-k)\}, \ldots, \min\{k, m\}\}$, and distribution $\mathbb{P}(\mathsf{x} = l) = \binom{k}{l}\binom{n-k}{m-l}/\binom{n}{m}$ for $l \in \mathcal{S}$. The expectation of x is $\mathbb{E}[\mathsf{x}] = \sum_{l\in\mathcal{S}} l\mathbb{P}(\mathsf{x} = l) = \frac{km}{n}$. Note this is the same as if the m balls were drawn *with* replacement, where $\mathsf{x} \sim \mathrm{Bin}(m, k/n)$. This expectation is pertinent in the derivation below of the expected degree when sampling an ER random graph using SS-C. Pick uniformly at random any vertex in G_t, i.e., a vertex that was not selected in the first t draws. This vertex has a random initial degree in G_0 of $\mathsf{d}_0 \sim \mathrm{Bin}(n-1, s)$. Let x_{t-1} be the random number of neighbors that are removed in the first $t-1$

samples, so that its random degree in G_t is $\mathsf{d}_t = \mathsf{d}_0 - \mathsf{x}_{t-1}$. Then:

$$\begin{aligned}
\mathbb{E}[\mathsf{d}_t] &= \mathbb{E}[\mathbb{E}[\mathsf{d}_t|\mathsf{d}_0]] \\
&= \mathbb{E}[\mathbb{E}[\mathsf{d}_0 - \mathsf{x}_{t-1}|\mathsf{d}_0]] \\
&= \mathbb{E}[\mathsf{d}_0 - \mathbb{E}[\mathsf{x}_{t-1}|\mathsf{d}_0]] \\
&= \mathbb{E}\left[\mathsf{d}_0 - \frac{\mathsf{d}_0(t-1)}{n-1}\right] \\
&= \left(1 - \frac{t-1}{n-1}\right)\mathbb{E}[\mathsf{d}_0] \\
&= \left(1 - \frac{t-1}{n-1}\right)(n-1)s \\
&= (n-1)s - (t-1)s = (n-t)s. \qquad (6.58)
\end{aligned}$$

The proof that $\mathbb{E}[\mathsf{x}_{t-1}|\mathsf{d}_0] = \frac{\mathsf{d}_0(t-1)}{n-1}$ comes from the discussion above, where the $n-1$ "balls" are the potential neighbors in G_0 of the randomly selected vertex, of which the d_0 marked "balls" are the actual neighbors, and $t-1$ "balls" are drawn.

Next, recall from the proof of Prop. 20 that

$$\mathbb{P}(\mathsf{c}_u \geq t|\mathsf{n}_0^{e,*}) = \prod_{u=1}^{t-1}\left(1 - \frac{\mathsf{n}_0^{e,*}}{n-u+1}\right), \qquad (6.59)$$

and observe the maximum number of samples possible under SS-C, denoted $\mathsf{c}_{\max}^{\mathrm{SS-C}}(\mathsf{n}_0^{e,*})$, is $n - \mathsf{n}_0^{e,*} + 1$. Recalling that $\mathbb{E}[\mathsf{z}_t] = 1 + \mathbb{E}[\mathsf{d}_t]$, substitution of these quantities into Eq. (6.54) yields Eq. (6.57). □

SS-S

As was the case with SS-C, leveraging the approximation in Eq. (6.54) for SS-S requires the expected extended degree of the random star center selected in sample t, as well as the distribution of the unit cost.

Proposition 19 (Linear cost of SS-S for ER graph.)**.** *The approximate expected linear cost under SS-S for an ER graph, conditioned on the random extended target set order* $\mathsf{n}_0^{e,*}$*, is* $\mathbb{E}[\tilde{\mathsf{c}}_l^{\mathrm{SS-S}}|\mathsf{n}_0^{e,*}] =$

$$\sum_{t=1}^{t_{\tilde{p}}^{(1)}}\left[1 + s\left((n-1)s\bar{s}^{t-1} - \frac{\bar{s}}{s}\left(1 - \bar{s}^{t-1}\right)\right)\right]\prod_{u=1}^{t-1}\left(1 - \tilde{p}_u^{\mathrm{SS-S}}\right), \qquad (6.60)$$

where $t_{\tilde{p}}^{(1)}$ *(c.f. Eq.* (6.31)*) and* $\tilde{p}_t^{\mathrm{SS-S}}$ *(c.f. Eq.* (6.29)*) are defined in Thm. 13.*

Proof. The proof follows the same lines as that of Prop. 18, i.e., first establish *i)* $\mathbb{E}[\mathsf{d}_t]$, *ii)* $\mathsf{c}_{\max}^{\mathrm{SS-S}}|\mathsf{n}_0^{e,*}$,

and *iii*) $\mathbb{P}(\mathsf{c}_u \geq t|\mathsf{n}_0^{e,*})$.

i) $\mathbb{E}[\mathsf{d}_t]$. The expected degree of the star center in sample t is obtained by a slight modification of the argument used in deriving $\mu_{t-1} \equiv \mathbb{E}[\mathsf{n}_{t-1}]$ in Cor. 4, given by Eq. (6.35). Specifically, consider any vertex not yet removed by the first $t-1$ samples. Conditioned on n_{t-1}, this vertex has a random degree $\mathsf{d}_t|\mathsf{n}_{t-1} \sim \text{Bin}(\mathsf{n}_{t-1}-1, s)$, and as such

$$\mathbb{E}[\mathsf{d}_t] = \mathbb{E}[\mathbb{E}[\mathsf{d}_t|\mathsf{n}_{t-1}]] = s\mathbb{E}[\mathsf{n}_{t-1}-1], \tag{6.61}$$

and as such, via Lem. 20 where the initial size of the watch set is $\mathsf{n}_{W,0} = n-1$ and the immune set size is $n_Z = 0$ it follows by Eq. (6.69) that

$$\mathbb{E}[\mathsf{d}_t] = s\left((n-1)\bar{s}^{t-1} - \frac{\bar{s}}{s}(1-\bar{s}^{t-1})\right). \tag{6.62}$$

ii) $c_{\max}^{\text{SS}-\text{S}}(\mathsf{n}_0^{e,*})$. The maximum number of samples before the target set is reached may be upper bounded by $t_{\tilde{p}}^{(1)}$ the time at which $\tilde{p}_{t+1}^{\text{SS}-\text{S}} = 1$ given in Eq. (6.31) of Thm. 13, i.e. $c_{\max}^{\text{SS}-\text{S}}(\mathsf{n}_0^{e,*}) \leq t_{\tilde{p}}^{(1)}$.

iii) $\mathbb{P}(\mathsf{c}_u \geq t|\mathsf{n}_0^{e,*})$. The probability of requiring t or more samples under SS-S is approximated by leveraging the results in Thm. 14, namely,

$$\mathbb{P}(\mathsf{c}_u \geq t|\mathsf{n}_0^{e,*}) \approx \prod_{u=1}^{t-1}(1-\tilde{p}_u^{\text{SS}-\text{S}}), \tag{6.63}$$

where $\tilde{p}_t^{\text{SS}-\text{S}}$ is defined in Eq. (6.29) in Thm. 13.

As in the case of Prop. 18, Eq. (6.60) follows by substitution. □

Numerical results for the estimates of the expected linear sampling cost of an ER graph under SS-R (Prop. 17), SS-C (Prop. 18), and SS-S (Prop. 19) sampling for Scenario *i*) and Scenario *ii*) are shown in Fig. 6.9 and Fig. 6.10. Under Scenario *i*), for small s, the estimates and the expected empirical linear costs diverge slightly, perhaps due to the underestimate of $\mathbb{E}[\mathsf{n}_0^{e,*}]$ shown in Fig. 6.2. Under Scenario *ii*) there is also a slight divergence between the estimated and the expected empirical linear cost for small s, but generally the expected cost estimates are more accurate then under Scenario *i*). Note that, under Scenario *i*) and *ii*), $\tilde{c}_l^{\text{SS}-\text{R}} \geq \tilde{c}_l^{\text{SS}-\text{S}}$ and $\tilde{c}_l^{\text{SS}-\text{C}} \geq \tilde{c}_l^{\text{SS}-\text{S}}$.

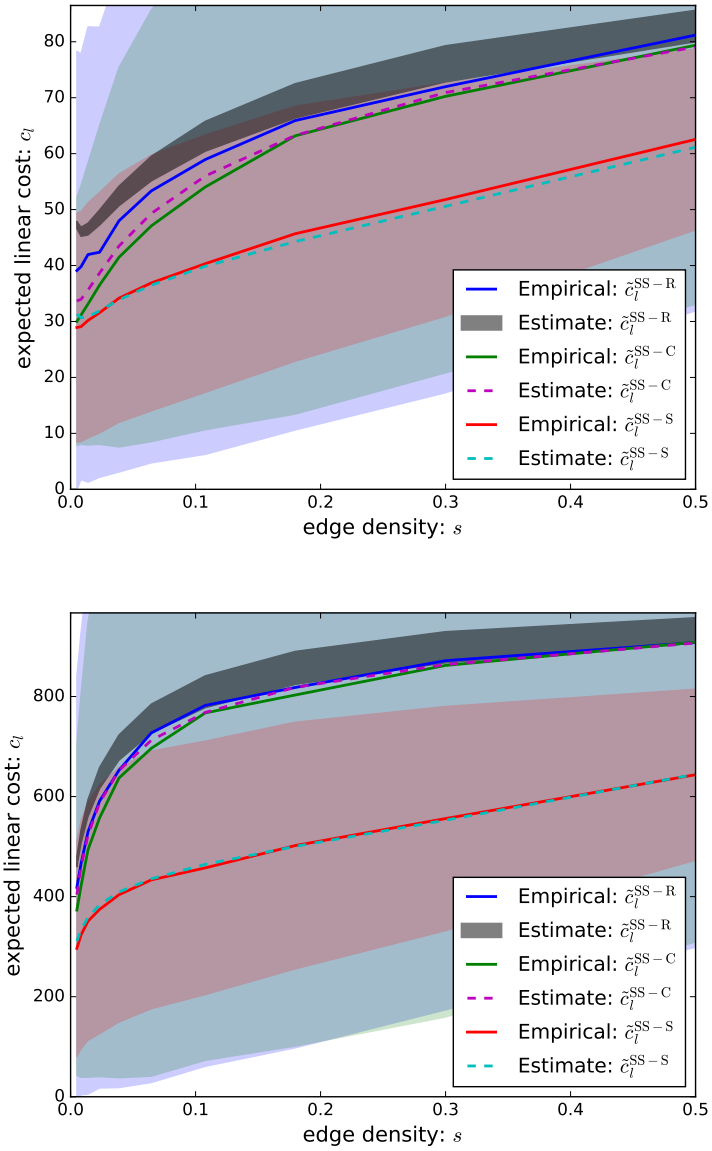

Figure 6.9: Expected linear cost for Scenario i) with graphs of order $n = 100$ (top) and $n = 1000$ (bottom): $s = 0.005$, 100 trials, 100 graphs.

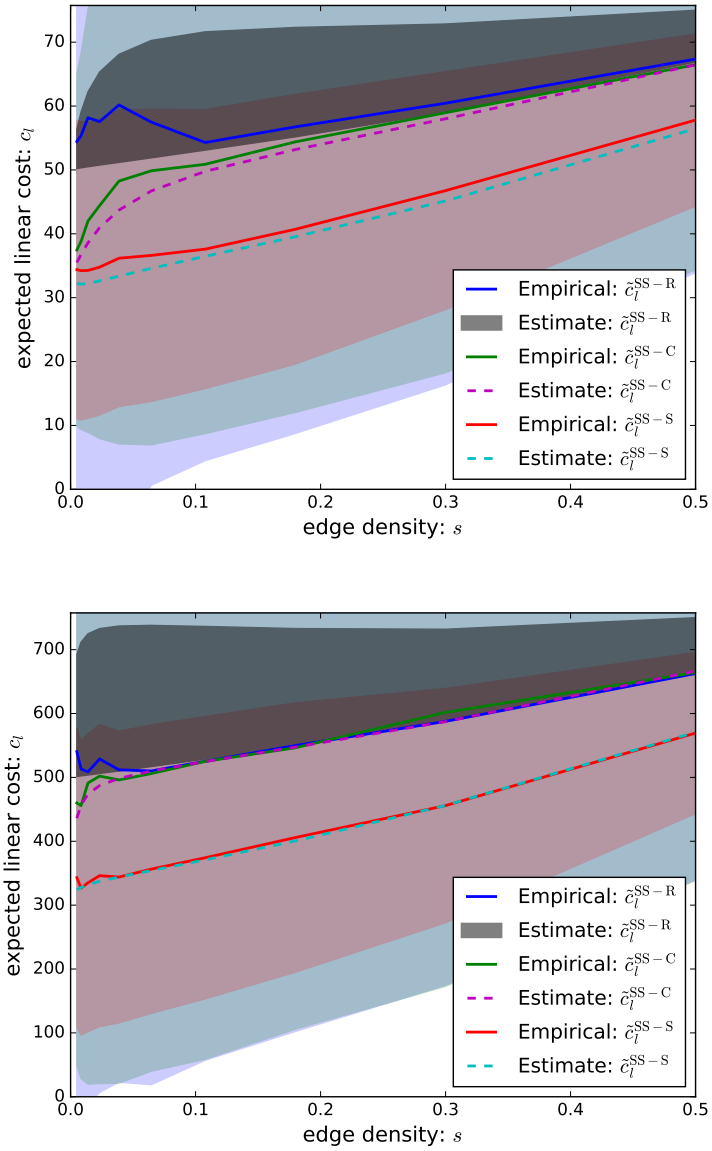

Figure 6.10: Expected linear cost for Scenario *ii*) with graphs of order $n = 100$ (top) and $n = 1000$ (bottom): $s = 0.005$, 100 trials, 100 graphs, $n_0^* = 2$.

6.6 Results on "real-world" graphs

This section tests whether the unit and linear cost estimates for SS-R, SS-C, and SS-S derived for ER graphs in Sec. 6.3 and Sec. 6.5 are accurate on "real-world" graphs. These graphs come from a variety of sources: `power-network` is a graph studied by Watts and Strogatz Watts and Strogatz [1998]; `CondMat`, `Gnutella08`, `Gnutella04`, `web-google`, `AstroPh`, `Epinions`, `fb-combined`, `oregon1`, and `brightkite` are from the SNAP repository Leskovec and Krevl [2014]; while the `web-edu` and `tech-routers` graphs are from the network data repository Rossi and Ahmed [2015].

Table 6.1: Graph Statistics

	n	m	s	α	JSD 1D	JSD 2D	degree range	$\lvert\mathcal{V}_\lambda\rvert$
AstroPh	18,772	198,110	1.1×10^{-3}	0.21	0.553	0.779	1, 504	1
brightkite	58,228	214,078	1.3×10^{-4}	0.11	0.426	0.701	1, 1134	1
CondMat	23,133	93,497	3.5×10^{-3}	0.14	0.258	0.508	1, 281	1
Epinions	75,879	405,740	1.4×10^{-4}	-0.04	0.653	0.891	1, 3044	1
fb-combined	4,039	88,234	1.1×10^{-2}	0.06	0.596	0.786	1, 1045	1
Gnutella04	10,876	39,994	6.8×10^{-4}	-0.01	0.356	0.628	1, 103	1
Gnutella08	6,301	20,777	1.0×10^{-3}	0.04	0.378	0.668	1, 97	1
oregon1	10,670	22,002	3.9×10^{-4}	-0.19	0.296	0.758	1, 2312	1
power-network	4,941	6,594	5.4×10^{-4}	0.00	0.058	0.184	1, 19	1
tech-routers	2,113	6,632	3.0×10^{-3}	0.02	0.344	0.642	1, 109	1
web-edu	3,031	6,474	1.4×10^{-3}	-0.17	0.422	0.823	1, 104	1
web-google	1,299	2,773	3.3×10^{-3}	-0.05	0.297	0.561	1, 59	1

Statistics for these graphs are given in Tab. 6.1, including order (n), size (m), edge density (s), and assortativity (α). Also included is the Jensen-Shannon divergence (JSD) between the graph's degree distribution and a binomial distribution (parametrized by n, s) (JSD 1D), and the JSD between the joint distribution of the degrees of an edge and a joint binomial distribution (parametrized by n, s) (JSD 2D). These measure in some way the "distance" between the graph and a corresponding ER random graph.

Empirical and estimated results for SS-R, SS-C, and SS-S sampling on a collection of real-world graphs assuming $\mathsf{n}_0^{e,*}$ is known is given for the unit cost model under Scenario i) and ii) in Tab. 6.2 and Tab. 6.3, and for the linear cost model in Tab. 6.4 and Tab. 6.5. Tab. 6.2 and Tab. 6.4 indicate that the expected unit and linear cost estimates for all three variates of star sampling under Scenario i) largely fall within the 95% confidence interval of the empirical trials and the SS-S cost estimates outside the 95% confidence intervals fall just beyond the intervals end points.

On the other hand Tab. 6.3 and Tab. 6.5 indicate that while expected unit and linear cost estimates for SS-R and SS-C sampling under scenario ii) fall within the 95% confidence interval, the estimates for the expected unit and linear cost of SS-S sampling under estimate the unit cost and over estimate the linear cost.

In particular looking at the absolute relative error, under scenario i) the relative error is low as is expected from the results above, however under scenario ii) the SS-S cost estimates in at least half the real-world graphs have at least 10% relative error. It is not clear why the SS-S cost estimates perform better in scenario i) then scenario ii), perhaps the ER assumptions made in estimating the expected cost under SS-S sampling become stronger as n_0^* increases or as degree of the nodes $v \in \mathcal{V}^*$ decrease.

Table 6.2: Unit cost scenario i), 1000 trials: Est. $\rightarrow$ Estimate, Con. $\rightarrow$ 95% Confidence Interval, Err. $\rightarrow$ Absolute Relative Error (%). Bold indicates the estimate is outside the confidence interval.

	SS-R			SS-C			SS-S		
	Est.	Con.	Err.	Est.	Con.	Err.	Est.	Con.	Err.
AstroPh	37.2	34.3, 40.2	0.2	37.1	34.5, 40.3	0.7	37.1	34.4, 40.5	0.9
britekite	51.3	46.1, 54.2	2.3	51.3	44.9, 52.9	4.8	51.3	48.9, 57.4	3.5
ContMat	82.0	74.3, 87.3	1.5	81.8	75.2, 88.0	0.2	81.7	77.5, 90.8	2.8
Epinions1	24.9	22.9, 27.0	0.26	24.9	23.4, 27.5	2.0	24.9	23.6, 27.6	2.8
fb-combined	3.9	3.5, 4.1	1.9	3.9	3.7, 4.3	3.0	3.9	3.5, 4.0	3.2
Gnutella04	104.6	97.8, 114.8	1.6	103.6	98.0, 115.4	2.9	103.4	100.3, 118.6	5.5
Gnutella08	64.3	60.4, 70.9	2.0	63.7	58.2, 68.8	0.25	63.6	60.1, 71.4	3.3
oregon1	4.6	4.5, 5.2	4.0	4.6	4.2, 4.9	1.3	4.6	4.1, 4.7	4.8
power-network	247.1	228.8, 268.0	0.5	235.3	217.4, 252.2	0.2	232.4	222.4, 256.0	2.8
tech-routers	19.2	17.6, 20.8	0.1	19.1	17.7, 20.7	0.7	19.0	18.0, 21.1	2.6
web-edu	28.9	27.4, 32.0	2.9	28.6	26.3, 30.9	0.0	**28.6**	32.3, 38.6	19.3
web-google	21.7	20.5, 24.2	3.1	21.3	18.7, 22.2	4.4	21.3	19.7, 22.9	0.0

Table 6.3: Unit cost scenario ii), 1000 trials $n_0^* = 4$: Est. $\rightarrow$ Estimate, Con. $\rightarrow$ 95% Confidence Interval, Err. $\rightarrow$ Absolute Relative Error (%). Bold indicates the estimate is outside the confidence interval.

	SS-R			SS-C			SS-S		
	Est.	Con.	Err.	Est.	Con.	Err.	Est.	Con.	Err.
AstroPh	86.6	76.1, 90.2	5.3	81.3	77.2, 89.7	0.4	78.3	69.6, 80.0	4.7
britekite	1164.6	1049.2, 1231.4	2.1	1141.7	1070.2, 1260.4	2.0	**1134.6**	1279.1, 1522.5	19.0
ContMat	545.2	547.4, 644.7	5.3	550.8	546.0, 643.5	7.4	545.0	508.6, 591.0	0.9
Epinions1	465.5	426.1, 497.8	1.4	465.5	418.2, 488.1	2.7	**465.1**	1020.5, 1298.3	60.0
fb-combined	28.0	26.0, 30.4	0.7	27.9	25.8, 30.2	0.5	27.7	28.9, 34.3	12.3
Gnutella04	302.1	268.7, 318.5	2.9	294.0	279.1, 326.2	2.9	290.4	288.9, 245.7	8.5
Gnutella08	525.1	479.5, 562.7	0.8	484.8	461.0, 537.5	2.9	**444.9**	459.0, 537.0	10.6
oregon1	304.9	273.4, 320.5	2.7	296.4	261.1, 308.0	4.2	**294.4**	343.0, 403.0	21.1
power-network	308.8	270.1, 319.1	4.8	290.7	266.9, 311.8	0.5	285.2	248.4, 287.4	6.4
tech-routers	192.0	176.6, 206.6	0.2	176.2	168.8, 196.5	3.6	**160.4**	226.9, 261.5	34.3
web-edu	216.5	214.1, 251.5	7.0	202.1	192.7, 223.4	2.8	**194.0**	226.0, 262.5	20.6
web-google	86.6	76.1, 90.2	4.1	81.3	77.2, 89.7	2.6	78.3	69.6, 80.0	4.7

Table 6.4: Linear cost scenario i), 1000 trials: Est. $\rightarrow$ Estimate, Con. $\rightarrow$ 95% Confidence Interval, Err. $\rightarrow$ Absolute Relative Error (%). Bold indicates the estimate is outside the confidence interval.

	SS-R			SS-C			SS-S		
	Est.	Con.	Err.	Est.	Con.	Err.	Est.	Con.	Err.
AstroPh	821.8	762.1, 884.2	0.2	818.6	762.5, 883.8	0.5	**788.1**	683.9, 786.9	7.2
britekite	428.5	385.1, 449.6	2.7	427.8	377.0, 441.6	4.5	425.5	380.6, 439.8	3.7
ContMat	745.1	675.8, 792.2	1.5	740.2	680.4, 794.3	0.4	722.1	656.2, 760.4	1.9
Epinions1	291.4	259.8, 307.6	2.7	291.2	271.5, 322.4	1.9	290.3	255.2, 301.7	4.3
fb-combined	172.6	156.4, 182.4	1.9	172.3	163.3, 190.9	2.7	167.2	144.5, 166.3	7.6
Gnutella04	873.7	816.8, 958.4	1.6	858.3	812.2, 954.8	2.9	808.7	729.2, 845.5	2.7
Gnutella08	488.3	460.1, 537.8	2.1	479.3	436.6, 512.6	1.0	**453.4**	387.4, 448.9	8.4
oregon1	23.6	19.0, 29.8	3.0	23.6	17.2, 32.4	4.8	23.6	17.6, 23.8	14.0
power-network	906.5	838.0, 981.7	0.4	834.9	774.2, 894.1	0.1	762.5	738.1, 839.4	3.3
tech-routers	139.8	127.0, 148.7	1.4	137.5	126.3, 146.2	0.9	**131.4**	113.5, 129.7	8.1
web-edu	152.2	143.8, 164.6	1.3	149.6	138.1, 158.5	0.9	145.1	145.1, 166.6	6.9
web-google	114.1	107.3, 126.9	2.6	110.8	97.5, 115.6	4.0	105.1	91.8, 105.4	6.6

Table 6.5: Linear cost scenario *ii*), 1000 trials, $n_0^* = 4$: Est. $\rightarrow$ Estimate, Con. $\rightarrow$ 95% Confidence Interval, Err. $\rightarrow$ Absolute Relative Error (%). Bold indicates the estimate is outside the confidence interval.

	SS-R			SS-C			SS-S		
	Est.	Con.	Err.	Est.	Con.	Err.	Est.	Con.	Err.
AstroPh	8829.6	7678.2, 9066.4	5.5	8477.6	7852.1, 9174.3	0.4	**5959.2**	4271.7, 4749.2	32.1
britekite	9727.7	8753.0, 10275.4	2.2	9375.6	8752.2, 10279.0	1.5	**8309.0**	6155.9, 7005.0	26.3
ContMat	5125.0	4974.1, 5858.9	5.4	4899.7	4849.9, 5693.8	7.1	**4174.0**	3435.9, 3901.4	13.5
Epinions1	5477.5	4981.9, 5782.3	1.8	5413.6	4885.4, 5669.5	2.6	5069.6	4944.2, 5765.5	4.8
fb-combined	1253.5	1165.7, 1361.4	0.8	1236.8	1147.9, 1334.7	0.4	**928.8**	816.5, 915.2	7.3
Gnutella04	2524.0	2245.2, 2660.5	2.9	2399.1	2278.6, 2652.2	2.7	**2034.3**	1738.3, 1980.4	9.4
Gnutella08	3987.9	3648.3, 4280.2	0.6	3453.3	3293.7, 3801.3	2.7	**2388.9**	1805.8, 2006.7	25.3
oregon1	1562.1	1388.9, 1632.3	3.4	1485.8	1320.1, 1545.9	3.7	**1354.9**	793.9, 902.8	59.7
power-network	1133.1	990.2, 1169.6	4.9	1023.5	941.8, 1094.7	0.5	912.5	812.1, 926.4	5.0
tech-routers	1397.9	1287.7, 1503.6	0.2	1196.8	1154.1, 1328.5	3.6	**819.7**	717.0, 792.0	8.6
web-edu	1141.4	1137.1, 1335.5	7.7	1011.6	972.0, 1117.5	3.2	815.6	725.3, 820.5	5.5
web-google	456.3	401.1, 475.1	4.2	407.7	388.0, 448.6	2.5	**330.8**	265.6, 296.3	17.7

6.7 Related work

Star sampling is presented as a special case of the more general concept of *snowball sampling* in Kolaczyk [2009]. *Snowball sampling* was introduced by Goodman [1961] and studied by Frank [1977]. *Snowball sampling* appears in Lee et al. [2006], Ahn et al. [2007], Hu and Lau [2013]. *Star sampling* is a snowball sample where a sample consists of a center vertex selected uniformly at random $\mathsf{v} \in \text{Uni}(\mathcal{V})$ and its immediate neighbors $\Gamma_v(\mathsf{v})$; *Star sampling* appears in Wang and Lu [2013].

This chapter is an extension of the work in Chap. 5 and Stokes and Weber [2017], both of which focused on estimating the number of star samples required to find a target vertex. The work in Chap. 5 considered the slightly different problem of finding a degree j vertex or an edge with an endpoint of degree j and degree k. Moreover, Chap. 5 attempted to analyze the performance of *star sampling with replacement* on a modified Erdős-Rényi (ER) random graph construction, where the endpoint degree of a randomly selected edge are truly independent.

There has been substantial work on graph sampling – far too much to credibly review here. Classic graph exploration strategies include: *random sampling* of vertices or edges, *random walk sampling*, and *random jump sampling*, which alternates between a *random walk* and *random sampling*. These graph exploration strategies are general in the sense that they perform reasonably well in a broad range of problems. However, the graph sampling literature itself is divided between work *i*) attempting to derive an unbiased or uniform estimate of the vertices in a network, and *ii*) attempting to find vertices with particular properties, for instance maximum degree vertices. The former problem is referred to as the *graph sampling problem* and the latter as the *graph search problem*.

The graph sampling problem became widespread with the advent of social media. In particular,

one important question it addresses is how to obtain a representative sample of social media users. To solve this problem Leskovac introduced *forest fire sampling* for temporal graphs where, similar to *breadth-first search* (BFS), a search frontier is established. However, instead of expanding this frontier to unexplored vertices as in BFS, each iteration there is a chance of the frontier retreating to re-examine previously explored vertices Leskovec and Faloutsos [2006]. Riberio proposed and analyzed a related algorithm entitled *frontier sampling* Ribeiro and Towsley [2010]. Avrachenkov et al. [2010] and Jin et al. [2011] have both proposed *random walk jump* algorithms to obtain an unbiased sample of vertices and Avin and Krishnamachari [2008] has proposed a *random walk* biased toward high degree unvisited vertices. Miaya has looked at the sampling bias of *degree biased random walks* showing that *expansion sampling* can be more effective means of exploring graphs Maiya and Berger-Wolf [2011]. Although subsequently Voudigari has proposed a *degree biased breadth-first search* algorithm for the graph sampling problem Voudigari et al. [2016].

More sophisticated algorithms for solving the graph sampling problem include *Metropolized random walk with backtracking*, proposed by Stutzbach et al. [2009]. Although Lee has argued that Metropolis-Hastings sampling algorithms should avoid backtracking Lee et al. [2012]. Li proposed a *Rejection controlled Metropolis-Hastings* algorithm and a *Non-backtracking generalized maximum-degree sampling* algorithm Li et al. [2015]. Gjorka found that *Metropolis-Hasting random walk*'s and *Re-weighted random walk*'s both out perform a simple random walk in returning a uniform sample of Facebook users Gjoka et al. [2010]. While Kurant et al. [2011] has shown that *weighted random walks* can be used to carry out stratified sampling on graphs, and Chiericetti et al. [2016] gives bounds on the number of steps required to return a uniform sample of a network using *rejection sampling*, *maximum-degree sampling*, and *Metropolis-Hastings sampling*.

The performance of a *random walk* in solving a graph search problems depends on its performance in the graph cover problem, the time it takes a random walk to visit every vertex $v \in \mathcal{V}$, or every vertex $v \in \mathcal{V}^*$ for $\mathcal{V}^* \subset \mathcal{V}$. This problem gained prominence with P2P networks where the question was how to design P2P networks and search algorithms which allowed users to efficiently locate files. Ikeda has shown that given any undirected connected graph G of order n the cover time and mean hitting time of a *degree biased random walk* is bounded by $O(n^2 \log n)$ and $O(n^2)$ respectively Ikeda and Kubo [2003]. Cooper has shown in sparse Erdős Rényi graphs $G(n, s)$ the cover time of a *random walk* is asymptotically $cn \log \frac{c}{c-1} \log n$ where $s = \frac{c \log n}{n}$ and $c > 1$ Cooper and Frieze [2007]. Cooper also shows that in power-law graphs of order n with parameter $c \geq 3$, finding all vertices of degree n^a or greater with a *degree biased random walk* gives $0 \leq a \leq 1$ and bias coefficient $b > 0$ is $\tilde{O}(n^{1-2ab(1-\epsilon)})$ with high probability Cooper et al. [2014].

Cooper's results match Adamic's observation that the search time of *random walks* and *degree biased random walks* scale sublinearly with the size of power-law graphs Adamic et al. [2001]. Similarly Lv et al. [2002] has also shown that for the graph search problem, random walks outperform network flooding in P2P networks; Gkantsidis et al. [2006] expanded on this work and Brautbar and Kearns [2010] and Avrachenkov et al. [2012] have both shown that *random walk jump* algorithms are effective in finding the high degree vertices. The work presented in Chap. 4 proposed a *self avoiding degree biased random walk jump* algorithm called SAWJ.

Random walks however are not the only approach to searching a graph for vertices with particular properties. Avrachenkov has introduced the *Two-stage algorithm* for finding high degree vertices developed under the assumption that queries of the sampled graph are limited Avrachenkov et al. [2014]. Given this assumption it has been shown in Chap. 2 and Chap. 3 that *biased random walks* and *star sampling* can both be effective in finding vertices of interest.

6.8 Conclusion

Star sampling is a natural graph sampling paradigm, and as such it is important to optimize its design. This chapter studied three star sampling variants, involving various types of replacement, motivated by analogous sampling strategies of balls from an urn. The analytical and simulation results demonstrate that the mathematical approximations lead to reasonably accurate performance estimators in both the unit and linear cost models on ER graphs, and this chapter proves there is no significant difference between the three variants in the unit cost model. In contrast to the unit cost model, numerical results suggest SS-S significantly outperforms SS-R and SS-C in the linear cost model. This chapter shows that the cost approximations given are reasonably accurate on "real-world" graphs.

6.9 Appendix A

Lem. 18 and Lem. 19 show that ER graphs are closed under SS-C and SS-S sampling, respectively.

Lemma 18. *For $t' \in \mathbb{N}$, take t' samples from an ER random graph G with parameters (n, s) using SS-C, yielding a sequence of random graphs $(\mathsf{G}_t, t \in [t'])$. Then $\mathsf{G}_t = (\mathsf{V}_t, \mathsf{E}_t)$ is an ER random graph with parameters $(n - t, s)$ for each $t \in [t']$.*

Proof. Let $\mathsf{G}_0 = \mathsf{G}$ denote the initial ER random graph, with $\mathsf{G}_0 = (\mathcal{V}_0, \mathsf{E}_0)$, where $\mathcal{V}_0 = [n]$ and E_0 may be considered as a sequence of $\binom{n}{2}$ IID Bernoulli RVs, say $\mathsf{E}_0 = (\mathsf{x}_e, e \in [\binom{n}{2}])$, with $\mathsf{x}_e \sim \text{Ber}(s)$ indicating the random inclusion or exclusion of an edge at "site" indexed by the unordered vertex

pair e.

Define the sequence of random vertex and set triples $((\mathsf{u}_t, \mathsf{V}_t, \mathsf{E}_t), t \in [t'])$, with $\mathsf{u}_t \sim \mathrm{Uni}(\mathsf{V}_{t-1})$ the randomly selected star center in sample t, $\mathsf{V}_t = \mathsf{V}_{t-1} \setminus \{\mathsf{u}_t\}$ the set of vertices that remain after removal of star center u_t, and $\mathsf{E}_t = \mathsf{E}_{t-1} \setminus \mathsf{F}_t$, for $\mathsf{F}_t \equiv \mathcal{N}_{\mathsf{u}_t}$, the edges that remain after removal of the edge neighborhood of u_t.

For any $t \in [t']$: *i)* $\mathsf{V}_t = \mathcal{V}_0 \setminus \mathsf{V}_t^c$, for $\mathsf{V}_t^c \equiv \mathsf{u}_1 \cup \cdots \cup \mathsf{u}_t$ the set of vertices removed in the first t samples, and *ii)* $\mathsf{E}_t = \mathsf{E}_0 \setminus \mathsf{E}_t^c$, for $\mathsf{E}_t^c \equiv \mathsf{F}_1 \cup \cdots \cup \mathsf{F}_t$ the set of edges removed in the first t samples. Then $|\mathsf{V}_t| = n - t$, and as E_0 is IID Ber(s), it follows that the RVs in E_t are IID Ber(s), so that G_t is an ER random graph (with random vertex labels) with parameters $(n - t, s)$. □

Lemma 19. *For $t' \in \mathbb{N}$, take t' samples from an ER random graph G with parameters (n, s) using SS-S, yielding a sequence of random graphs $(\mathsf{G}_t, t \in [t'])$. Then $\mathsf{G}_t = (\mathsf{V}_t, \mathsf{E}_t)$ is an ER random graph with random order parameter n_t and edge probability s, for each $t \in [t']$. The random order parameter, $\mathsf{n}_t \equiv |\mathsf{V}_t|$, obeys the recursion $\mathsf{n}_t = \mathsf{n}_{t-1} - \mathsf{d}_t^e$, for $\mathsf{d}_t^e \sim 1 + \mathrm{Bin}(\mathsf{n}_{t-1} - 1, s)$, for $t \in [t']$.*

Proof. The first paragraph in the proof of Lem. 18 holds, and assume the notation therein.

Define the sequence of random vertex and set triples $((\mathsf{u}_t, \mathsf{V}_t, \mathsf{E}_t), t \in [t'])$, with $\mathsf{u}_t \sim \mathrm{Uni}(\mathsf{V}_{t-1})$ the randomly selected star center in sample t, $\mathsf{V}_t = \mathsf{V}_{t-1} \setminus \Gamma^e_{\mathsf{u}_t}$ the set of vertices that remain after removal of *star* $\Gamma^e_{\mathsf{u}_t}$, and $\mathsf{E}_t = \mathsf{E}_{t-1} \setminus \mathcal{N}^e_{\mathsf{u}_t}$, the edges that remain after removal of the *extended* edge neighborhood of u_t.

The sequence of pairs of RVs $((\mathsf{n}_t, \mathsf{d}_t^e), t \in [t'])$, where $\mathsf{n}_t = |\mathsf{V}_t|$ is the random order of G_t and $\mathsf{d}_t^e = |\Gamma^e_{\mathsf{u}_t}|$ is the random order of the extended neighborhood of u_t, obeys (by construction) the recursion $\mathsf{n}_t = \mathsf{n}_{t-1} - \mathsf{d}_t^e$, with $\mathsf{n}_0 = n$.

Now proving by induction in t that *i)* G_t is ER, and *ii)* $\mathsf{d}_t^e \sim 1 + \mathrm{Bin}(\mathsf{n}_{t-1} - 1, s)$, for $t \in [t']$.

Base case. It must be shown *i)* $\mathsf{G}_1 = (\mathsf{V}_1, \mathsf{E}_1)$ is an ER random graph with parameters $\mathsf{n}_1 = n - \mathsf{d}_1^e$ and s, and *ii)* $\mathsf{d}_1^e \sim 1 + \mathrm{Bin}(n - 1, s)$. Recall $\mathsf{V}_1 = \mathcal{V}_0 \setminus \Gamma^e_{\mathsf{G}_0}(\mathsf{u}_1)$ and $\mathsf{E}_1 = \mathsf{E}_0 \setminus \mathcal{N}^e_{\mathsf{u}_1}$. First, the edges in E_1 are a (randomly selected) subset of the IID Ber(s) RVs in E_0, and as such the edges in E_1 are also IID Ber(s) RVs; this shows *i)*. Second, $\mathsf{d}_1^e = |\Gamma^e_{\mathsf{u}_1}|$ counts both the star center u_1 and its neighbors Γ_{u_1}, where $d_{\mathsf{u}_1} \equiv |\Gamma_{\mathsf{u}_1}|$ has distribution Ber$(n - 1, s)$ by construction; this shows *ii)*.

Induction hypothesis. Fix $t \in [t']$ and assume the induction hypothesis for $t - 1$, i.e., assume *i)* G_{t-1} is an ER random graph with parameters n_{t-1} and s, and *ii)* $\mathsf{d}_{t-1}^e \sim 1 + \mathrm{Bin}(\mathsf{n}_{t-2} - 1, s)$. It follows that $\mathsf{n}_{t-1} = \mathsf{n}_{t-2} - \mathsf{d}_{t-1}^e$. It must shown that this implies *i)* the random graph G_t obtained by removing a random star from G_{t-1} is an ER graph with parameters $\mathsf{n}_t = \mathsf{n}_{t-1} - \mathsf{d}_t^e$ and s, and *ii)* $\mathsf{d}_t^e \sim 1 + \mathrm{Bin}(\mathsf{n}_{t-1}, s)$. First, the edges in E_t are a (randomly selected) subset of the IID Ber(s) RVs

in E_{t-1}, and as such the edges in E_t are also IID Ber(s) RVs; this shows i). Second, $\mathsf{d}_t^e = |\Gamma_{\mathsf{u}_t}^e|$ counts both the star center u_t and its neighbors Γ_{u_t}, where $d_{\mathsf{u}_t} \equiv |\Gamma_{\mathsf{u}_t}|$ has distribution Ber($\mathsf{n}_{t-1} - 1, s$) by construction; this shows ii). □

6.10 Appendix B

This appendix holds Prop. 20 and its proof. This result is certainly already known, although being unaware of a reference, it has been included here for completeness.

Fix $n \in \mathbb{N}$ and $n^* \in [n]$. Consider sampling without replacement from an urn initially holding n balls of which n^* are marked, where the sampling procedure terminates upon the first draw of a marked ball.

Proposition 20. *The average number of samples (without replacement) from an urn with n balls, until one the n^* marked balls is selected, is $(n+1)/(n^*+1)$.*

Proof. Define:

- $(n)_k \equiv (n-0)(n-1)\cdots(n-(k-1))$ as the falling factorial (recall $\binom{n}{k} = (n)_k/k!$);
- $p_t \equiv n^*/(n-t+1)$ for $t \in [n-n^*+1]$ as the probability of success on trial t, conditioned on failure in the first $t-1$ trials (observe $p_1 = n^*/n$ and $p_{n-n^*+1} = 1$);
- $q_t \equiv p_t \prod_{c=1}^{t-1}(1-p_c)$ as the unconditioned probability of a first success on trial t;
- $\mathsf{x} \sim \mathbf{q}$ as the RV for the number of required trials, where $\mathbf{q} \equiv (q_t, t \in [n-n^*+1])$.

Nest its shown that $\mathbf{q}$ is normalized and that $\mathbb{E}[\mathsf{x}] = (n+1)/(n^*+1)$. Let $\Sigma_{\mathbf{q}} \equiv \sum_{t=1}^{n-n^*+1} q_t$.

Proof that $\mathbf{q}$ is correctly normalized.

$$
\begin{aligned}
\Sigma_{\mathbf{q}} &= \sum_{t=1}^{n+1-n^*} p_t \prod_{s=1}^{t-1}(1-p_s) \\
&= \frac{n^*}{n} + \sum_{t=2}^{n+1-n^*} \frac{n^*}{n-t+1} \prod_{s=1}^{t-1}\left(1-\frac{n^*}{n+1-s}\right) \\
&= \frac{n^*}{n} + \sum_{t=2}^{n+1-n^*} \frac{n^*}{n+1-t} \prod_{s=1}^{t-1}\frac{n+1-n^*-s}{n+1-s} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{(n-n^*)\cdots(n-n^*-(t-2))}{(n)\cdots(n-(t-2))(n-(t-1))} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{(n-n^*)_{t-1}}{(n)_t} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{(t-1)!\binom{n-n^*}{t-1}}{t!\binom{n}{t}} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{\binom{n-n^*}{t-1}}{t\binom{n}{t}} \\
&= \frac{n^*}{n} + \frac{n^*}{n+1-n^*} \sum_{t=2}^{n+1-n^*} \frac{\frac{n-n^*+1}{t}\binom{n-n^*}{t-1}}{\binom{n}{t}} \\
&= \frac{n^*}{n} + \frac{n^*}{n+1-n^*} \sum_{t=2}^{n+1-n^*} \frac{\binom{n+1-n^*}{t}}{\binom{n}{t}} \\
&= \frac{n^*}{n} + \frac{n^*}{n+1-n^*} \sum_{t=2}^{n+1-n^*} \frac{(n+1-n^*)!t!(n-t)!}{t!(n+1-n^*-t)!n!} \\
&= \frac{n^*}{n} + \frac{n^*(n-n^*)!}{n!} \sum_{t=2}^{n+1-n^*} \frac{(n-t)!}{(n+1-n^*-t)!}.
\end{aligned} \tag{6.64}
$$

Define $s = n+1-n^*-t$, and observe the sum ranges over s from 0 to $n-1-n^*$:

$$
\begin{aligned}
\Sigma_{\mathbf{q}} &= \frac{n^*}{n} + \frac{n^*(n-n^*)!}{n!} \sum_{s=0}^{n-1-n^*} \frac{(n^*-1+s)!}{s!} \\
&= \frac{n^*}{n} + \frac{n^*!(n-n^*)!}{n!} \sum_{s=0}^{n-1-n^*} \frac{(n^*-1+s)!}{s!(n^*-1)!} \\
&= \frac{n^*}{n} + \frac{1}{\binom{n}{n^*}} \sum_{s=0}^{n-1-n^*} \binom{n^*-1+s}{s} \\
&= \frac{n^*}{n} + \frac{1}{\binom{n}{n^*}} \binom{n-1}{n-1-n^*} \\
&= \frac{n^*}{n} + \frac{1}{\binom{n}{n^*}} \binom{n-1}{n^*} \\
&= \frac{n^*}{n} + \frac{(n-n^*)}{n} = 1.
\end{aligned} \tag{6.65}
$$

Proof that $\mathbb{E}[\mathsf{x}] = (n+1)/(n^*+1)$.

$$\begin{aligned}
\mathbb{E}[\mathsf{x}] &= \sum_{t=1}^{n+1-n^*} t q_t \\
&= p_1 + \sum_{t=2}^{n+1-n^*} t p_t \prod_{s=1}^{t-1}(1-p_s) \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{t}{n+1-t} \prod_{s=1}^{t-1}\left(1 - \frac{n^*}{n+1-s}\right) \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{t}{n+1-t} \prod_{s=1}^{t-1}\frac{n+1-n^*-s}{n+1-s} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} t\frac{(n-n^*)_{t-1}}{(n)_t} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} t\frac{(t-1)!\binom{n-n^*}{t-1}}{t!\binom{n}{t}} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{\binom{n-n^*}{t-1}}{\binom{n}{t}} \\
&= \frac{n^*}{n} + n^* \sum_{t=2}^{n+1-n^*} \frac{(n-n^*)!}{(t-1)!(n+1-n^*-t)!}\frac{t!(n-t)!}{n!} \\
&= \frac{n^*}{n} + \frac{n^*(n-n^*)!}{n!} \sum_{t=2}^{n+1-n^*} \frac{t(n-t)!}{(n+1-n^*-t)!}. \qquad (6.66)
\end{aligned}$$

Continuing, with $s = n+1-n^*-t$:

$$\begin{aligned}
\mathbb{E}[\mathsf{x}] &= \frac{n^*}{n} + \frac{n^*(n-n^*)!}{n!} \\
&\quad \times \sum_{s=0}^{n-1-n^*} \frac{(n+1-n^*-s)(n^*-1+s)!}{s!} \\
&= \frac{n^*}{n} + \frac{n^*!(n-n^*)!}{n!} \\
&\quad \times \sum_{s=0}^{n-1-n^*} \frac{(n+1-n^*-s)(n^*-1+s)!}{(n^*-1)!s!} \\
&= \frac{n^*}{n} + \frac{1}{\binom{n}{n^*}} \sum_{s=0}^{n-1-n^*} (n+1-n^*-s)\binom{n^*-1+s}{n^*-1}.
\end{aligned}$$

It is straightforward to verify that

$$\sum_{s=0}^{n-1-n^*} (n+1-n^*-s)\binom{n^*-1+s}{n^*-1} = \frac{(n+n^*+1)(n-1)!}{(n-n^*-1)!(n^*+1)!}. \qquad (6.67)$$

This yields:

$$\begin{aligned}\mathbb{E}[\mathsf{x}] &= \frac{n^*}{n} + \frac{1}{\binom{n}{n^*}} \times \frac{(n+n^*+1)(n-1)!}{(n-n^*-1)!(n^*+1)!} \\ &= \frac{n^*}{n} + \frac{(n-n^*)(n+n^*+1)}{n(n^*+1)} = \frac{n+1}{n^*+1}. \end{aligned} \tag{6.68}$$

□

6.11 Appendix C

This appendix establishes Lem. 20, the proof of Thm. 14, and the proof of Prop. 16.

Notation for Lem. 20. Let $(\mathsf{G}_t, t \in \mathbb{Z}^+)$ be a sequence of ER random graphs induced by SS-S, with $\mathsf{G}_t = (\mathsf{V}_t, \mathsf{E}_t)$ the graph following the removal of the random star selected in sample t, starting from an initial ER random graph $\mathsf{G}_0 = (\mathcal{V}_0, \mathsf{E}_0)$ with parameters (n, s). Set $\bar{s} \equiv 1 - s$.

Fix a "watch" subset $\mathcal{W}_0 \subseteq \mathcal{V}_0$, a "draw" subset $D_0 \subseteq \mathcal{V}_0$, and an "immune" subset $\mathcal{Z}_0 \subseteq \mathcal{W}_0$, those vertices in $\mathcal{W}_0$ with no neighbors in draw set $\mathcal{D}_0$. Define the random set-valued sequences $(\mathsf{W}_t, t \in \mathbb{Z}^+)$ and $(\mathsf{D}_t, t \in \mathbb{Z}^+)$, with $\mathsf{W}_t \equiv W_0 \cap \mathsf{V}_t$ and $\mathsf{D}_t \equiv \mathcal{D}_0 \cap \mathsf{V}_t$ the watch and draw sets surviving the first t samples. Two cases are considered: case i) the star center is *never* drawn from the watch set, i.e., $\mathcal{D}_0 \cap \mathcal{W}_0 = \emptyset$, and case ii) the star center is *always* drawn from the watch set, i.e., $\mathcal{D}_0 \subseteq \mathcal{W}_0$.

Let $(\mathsf{v}_t, t \in \mathbb{N})$, with $\mathsf{v}_t \sim \text{Uni}(\mathsf{V}_{t-1})$, denote the sequence of random star centers associated with each random sample, and define the event sequence $(\mathcal{C}_{\mathcal{D},t}, t \in \mathbb{N})$ that star center t is drawn from the surviving draw set, i.e., $\mathcal{C}_{\mathcal{D},t} \equiv \{\mathsf{v}_t \in \mathsf{D}_{t-1}\}$. Let $\bar{\mathcal{C}}_{\mathcal{D},t} \equiv \bigcap_{t' \in [t]} \mathcal{C}_{\mathcal{D},t'}$ be the event that the first t star centers are each drawn from the surviving draw set.

Define the random sequence $(\mathsf{n}_{\mathcal{W},t}, t \in \mathbb{Z}^+)$, with $\mathsf{n}_{\mathcal{W},t} \equiv |\mathsf{W}_t|$ the order of the watch set, and define the sequences of conditional means $(\mu_{\mathcal{W}|\mathcal{D},t}, t \in \mathbb{Z}^+)$ and conditional variances $(\sigma^2_{\mathcal{W}|\mathcal{D},t}, t \in \mathbb{Z}^+)$, with $\mu_{\mathcal{W}|\mathcal{D},t} \equiv \mathbb{E}[\mathsf{n}_{\mathcal{W},t}|\bar{\mathcal{C}}_{\mathcal{D},t}]$ and $\sigma^2_{\mathcal{W},t} \equiv \text{Var}(\mathsf{n}_{\mathcal{W},t}|\bar{\mathcal{C}}_{\mathcal{D},t})$ associated with $\mathsf{n}_{\mathcal{W},t}$ and $\bar{\mathcal{C}}_{\mathcal{D},t}$. Additionally let n_Z be the order of set $\mathcal{Z}_0$.

Lemma 20. *Under the notation above, the mean and variance of the size of the watch set, conditioned on drawing from the draw set, after t samples are given as follows.*

Case i) ($\mathcal{D}_0 \cap \mathcal{W}_0 = \emptyset$):

$$\mu^{i)}_{\mathcal{W}|\mathcal{D},t} = (n_{\mathcal{W},0} - n_Z)\bar{s}^t + n_{\mathcal{Z}} \tag{6.69}$$

$$\sigma^{2,i)}_{\mathcal{W}|\mathcal{D},t} = (n_{\mathcal{W},0} - n_{\mathcal{Z}})\bar{s}^t(1 - \bar{s}^t). \tag{6.70}$$

Case ii) ($\mathcal{D}_0 \subseteq \mathcal{W}_0$ with $n_{\mathcal{Z}} = 0$):

$$\mu^{ii)}_{\mathcal{W}|\mathcal{D},t} = \mu^{i)}_{\mathcal{W}|\mathcal{D},t} - \frac{\bar{s}}{s}(1-\bar{s}^t) \tag{6.71}$$

$$\sigma^{2,ii)}_{\mathcal{W}|\mathcal{D},t} = \sigma^{2,i)}_{\mathcal{W}|\mathcal{D},t} + \frac{\bar{s}}{s(1+\bar{s})}\left(1+\bar{s}(1-\bar{s}^t)\right)\bar{s}^t. \tag{6.72}$$

Proof. Let $\mathsf{r}_{\mathcal{W},t}$ be the RV denoting the number of vertices in the watch set *removed* by sample t. The random recurrence induced by the sampling is $\mathsf{n}_{\mathcal{W},t} = \mathsf{n}_{\mathcal{W},t-1} - \mathsf{r}_{\mathcal{W},t}$. First:

$$\begin{aligned}
\mu_{\mathcal{W}|\mathcal{D},t} &\equiv \mathbb{E}[\mathsf{n}_{\mathcal{W},t}|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \mathbb{E}[\mathbb{E}[\mathsf{n}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \mathbb{E}[\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - \mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - \mathbb{E}[\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}].
\end{aligned}$$

Second:

$$\begin{aligned}
\sigma^2_{\mathcal{W}|\mathcal{D},t} &\equiv \mathrm{Var}(\mathsf{n}_{\mathcal{W},t}|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&= \mathrm{Var}(\mathbb{E}[\mathsf{n}_{\mathcal{W},t}|\mathsf{n}_{t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}) + \mathbb{E}[\mathrm{var}(\mathsf{n}_{\mathcal{W},t}|\mathsf{n}_{t-1})|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \mathrm{Var}(\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - \mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&+ \mathbb{E}[\mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1} - \mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1})|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1} - \mathbb{E}[\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}]|\bar{\mathcal{C}}_{\mathcal{D},t}). \\
&+ \mathbb{E}[\mathrm{Var}(\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1})|\bar{\mathcal{C}}_{\mathcal{D},t}]
\end{aligned}$$

Consider case i) ($\mathcal{D}_0 \cap \mathcal{W}_0 = \emptyset$, $\mathcal{Z}_0 = \emptyset$). As the star center is *never* drawn from the watch set $\mathsf{r}_{\mathcal{W},t}|(\mathsf{n}_{\mathcal{W},t-1}, \mathcal{C}_{\mathcal{D},t}) \sim \mathrm{Bin}(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}, s)$, and therefore i) $\mathbb{E}[r_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}, \mathcal{C}_{\mathcal{D},t}] = (\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})s$ and ii) $\mathrm{Var}(\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}, \mathcal{C}_{\mathcal{D},t}) = (\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})s\bar{s}$. First:

$$\begin{aligned}
\mu_{\mathcal{W}|\mathcal{D},t} - n_{\mathcal{Z}} &= \mathbb{E}[(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}) - (\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})s|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
\mu_{\mathcal{W}|\mathcal{D},t} - n_{\mathcal{Z}} &= \bar{s}\mathbb{E}[(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
\mu_{\mathcal{W}|\mathcal{D},t} - n_{\mathcal{Z}} &= \bar{s}\mathbb{E}[(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})|\bar{\mathcal{C}}_{\mathcal{D},t-1}] \\
\mu_{\mathcal{W}|\mathcal{D},t} - n_{\mathcal{Z}} &= \bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - n_{\mathcal{Z}}).
\end{aligned}$$

The recurrence $\mu_{\mathcal{W}|\mathcal{D},t} - n_{\mathcal{Z}} = \bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - n_{\mathcal{Z}})$ with initial condition $\mu_{\mathcal{W}|\mathcal{D},0} = n_{\mathcal{W},0}$ has solution

Eq. (6.69).

$$\begin{aligned}
\sigma^2_{\mathcal{W}|\mathcal{D},t} &= \mathrm{Var}((\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}) - (\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})s|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&\quad + \mathbb{E}[(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}})s\bar{s}|\bar{\mathcal{C}}_{\mathcal{D},t-1}] \\
&= s^2\mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}|\bar{\mathcal{C}}_{\mathcal{D},t}) + s\bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}|\bar{\mathcal{C}}_{\mathcal{D},t-1}] \\
&= s^2\mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1}|\bar{\mathcal{C}}_{\mathcal{D},t-1}) + s\bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}|\bar{\mathcal{C}}_{\mathcal{D},t-1}] \\
&= s^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - n_{\mathcal{Z}}|\bar{\mathcal{C}}_{\mathcal{D},t-1}] \\
&= s^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - n_{\mathcal{Z}}) \\
&= s^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}((n_{\mathcal{W},0} - n_{\mathcal{Z}})\bar{s}^{t-1}) \\
&= s^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}^t(n_{\mathcal{W},0} - n_{\mathcal{Z}}).
\end{aligned}$$

The recurrence $\sigma^2_{\mathcal{W}|\mathcal{D},t} = \bar{s}^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}^t(n_{\mathcal{W},0} - n_{\mathcal{Z}})$ with initial condition $\sigma^2_{\mathcal{W}|\mathcal{D},0} = 0$ has solution Eq. (6.70).

Consider case *ii*) ($\mathcal{D}_0 \subseteq \mathcal{W}_0$, $\mathcal{Z}_0 = \emptyset$). As the star center is *always* drawn from the watch set and $n_{\mathcal{Z}} = 0$ $\mathsf{r}_{\mathcal{W},t}|(\mathsf{n}_{\mathcal{W},t-1}, \mathcal{E}_{\mathcal{D},t}) \sim 1 + \mathrm{Bin}(\mathsf{n}_{\mathcal{W},t-1} - 1, s)$, and therefore *i*) $\mathbb{E}[\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}] = 1 + (\mathsf{n}_{\mathcal{W},t-1} - 1)s$ and *ii*) $\mathrm{Var}(\mathsf{r}_{\mathcal{W},t}|\mathsf{n}_{\mathcal{W},t-1}) = (\mathsf{n}_{\mathcal{W},t-1} - 1)s\bar{s}$. First:

$$\begin{aligned}
\mu_{\mathcal{W}|\mathcal{D},t} &= \mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - (1 + (\mathsf{n}_{\mathcal{W},t-1} - 1)s)|\bar{\mathcal{C}}_{\mathcal{D},t}] \\
&= \bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1}|\bar{\mathcal{C}}_{\mathcal{D},t}] - \bar{s} \\
&= \bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1}|\bar{\mathcal{C}}_{\mathcal{D},t-1}] - \bar{s} \\
&= \bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - 1).
\end{aligned}$$

The recurrence $\mu_{\mathcal{W}|\mathcal{D},t} = \bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - 1)$ with initial condition $\mu_{\mathcal{W}|\mathcal{D},0} = n_{\mathcal{W},0}$ has solution Eq. (6.71).

Second:

$$\begin{aligned}
\sigma^2_{\mathcal{W}|\mathcal{D},t} &= \mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1} - (1 + (\mathsf{n}_{\mathcal{W},t-1} - 1)s)|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&\quad + \mathbb{E}[(\mathsf{n}_{\mathcal{W},t-1} - 1)s\bar{s}|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&= \bar{s}^2\mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1}|\bar{\mathcal{C}}_{\mathcal{D},t}) + s\bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - 1|\bar{\mathcal{C}}_{\mathcal{D},t}) \\
&= \bar{s}^2\mathrm{Var}(\mathsf{n}_{\mathcal{W},t-1}|\bar{\mathcal{C}}_{\mathcal{D},t-1}) + s\bar{s}\mathbb{E}[\mathsf{n}_{\mathcal{W},t-1} - 1|\bar{\mathcal{C}}_{\mathcal{D},t-1}) \\
&= \bar{s}^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\bar{s}(\mu_{\mathcal{W}|\mathcal{D},t-1} - 1) \\
&= \bar{s}^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + s\mu_{\mathcal{W}|\mathcal{D},t} \\
&= \bar{s}^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + \bar{s}\left(((n_{\mathcal{W},0} - 1)s + 1)\bar{s}^{t-1} - 1\right).
\end{aligned}$$

The recurrence

$$\sigma^2_{\mathcal{W}|\mathcal{D},t} = \bar{s}^2\sigma^2_{\mathcal{W}|\mathcal{D},t-1} + ((n_{\mathcal{W},0} - 1)s + 1)\bar{s}^t - \bar{s},$$

with initial condition $\sigma^2_{\mathcal{W}|\mathcal{D},0} = 0$ has solution Eq. (6.72). □

Proof of Thm. 14. First Eq. (6.43) is proved, then a proof of Eq. (6.44) is given.

Proof of Eq. (6.43). Set $\mathcal{C}_t$ as the event that star t misses the target, $\bar{\mathcal{C}}_t = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_t$ as the event that the first t stars each miss the target, and $\mathcal{C}_t^c$ as the complement of $\mathcal{C}_t$, i.e., the event that star t hits the target. Set $p_{t+1} \equiv \mathbb{P}(\mathcal{C}_{t+1}^c|\bar{\mathcal{C}}_t)$ as the *conditional* probability that star $t+1$ hits the target *given* that the first t stars miss the target. Note $1 - p_{t+1} = \mathbb{P}(\mathcal{C}_{t+1}|\bar{\mathcal{C}}_t)$. Set $q_{t+1} \equiv \mathbb{P}(\mathcal{C}_{t+1}^c \cap \bar{\mathcal{C}}_t)$ as the *unconditioned* probability that star $t+1$ hits the target *and* that the first t stars miss the target. Thus q_{t+1} is the unconditioned probability that the first star to hit the target is star $t+1$. Set $r_t \equiv \mathbb{P}(\bar{\mathcal{C}}_t)$ as the probability that the first t stars each miss the target. As

$$q_{t+1} = \mathbb{P}(\mathcal{C}_{t+1}^c|\bar{\mathcal{C}}_t)\mathbb{P}(\bar{\mathcal{C}}_t) = p_{t+1}r_t. \tag{6.73}$$

and

$$r_t = \mathbb{P}(\mathcal{C}_1)\mathbb{P}(\mathcal{C}_2|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_3|\bar{\mathcal{C}}_2)\cdots\mathbb{P}(\mathcal{C}_t|\bar{\mathcal{C}}_{t-1}) = \prod_{s=1}^{t}(1 - p_s), \tag{6.74}$$

it follows that $q_{t+1} = p_{t+1}\prod_{s=1}^{t}(1 - p_s)$. Given an approximation $\tilde{p}_{t+1} \approx p_{t+1}$ for the conditional distribution, this section obtains Eq. (6.43) as an approximation for the *unconditional* distribution.

Proof of Eq. (6.44). Consider a sequence $\hat{p} \equiv (\hat{p}_t, t \in \mathbb{N})$ with each $\hat{p}_t \in (0, 1)$. For $T \in \mathbb{N}$ set

$p^{(T)} \equiv (p_t^{(T)}, t \in [T])$, with

$$p_t^{(T)} \equiv \begin{cases} \hat{p}_t, & t \in [T-1] \\ 1, & t = T \end{cases}, \tag{6.75}$$

Define *i)* $\bar{p}_t^{(T)} \equiv 1 - p_t^{(T)}$; *ii)* $\pi^{(T)} \equiv (\pi_t^{(T)}, t \in [T])$, with $\pi_1^{(T)} \equiv 1$, $\pi_t^{(T)} \equiv \prod_{s\in[t-1]} \bar{p}_s^{(T)}$ for $t \in \{2, \ldots, T\}$; *iii)* $q^{(T)} \equiv (q_t^{(T)}, t \in [T])$, with $q_t^{(T)} \equiv p_t^{(T)} \pi_t^{(T)}$, noting $q_1^{(T)} = p_1^{(T)}$ and $q_T^{(T)} = \pi_T^{(T)}$; and *iv)* $\mu^{(T)} \equiv \sum_{t\in[T]} \pi_t^{(T)}$. Then: *a)* $q^{(T)}$ is a probability distribution, *b)* if $\mathsf{x}^{(T)} \sim q^{(T)}$ then $\mathbb{P}(\mathsf{x}^{(T)} \geq t) = \pi_t^{(T)}$, and *c)* $\mathbb{E}[\mathsf{x}^{(T)}] = \mu^{(T)}$.

a) Set $\Sigma_q^{(T)} \equiv \sum_{t\in[T]} q_t^{(T)}$. It must be shown that $\Sigma_q^{(T)} = 1$. The proof is by induction in T. The base case $T = 1$ is trivial. Suppose it is true for T, i.e., suppose $\Sigma_q^{(T)} = 1$, so that

$$\Sigma_q^{(T)} = \Sigma_q^{(T-1)} + \pi_T^{(T)} = 1. \tag{6.76}$$

Consider the induction hypothesis case $T + 1$:

$$\begin{aligned} \Sigma_q^{(T+1)} &= \sum_{t\in[T-1]} q_t^{(T+1)} + q_T^{(T+1)} + q_{T+1}^{(T+1)} \\ &= \sum_{t\in[T-1]} q_t^{(T+1)} + p_T^{(T+1)} \pi_T^{(T+1)} + \pi_{T+1}^{(T+1)} \\ &= \sum_{t\in[T-1]} q_t^{(T+1)} + p_T^{(T+1)} \pi_T^{(T+1)} + \bar{p}_T^{(T+1)} \pi_T^{(T+1)} \\ &= \sum_{t\in[T-1]} q_t^{(T+1)} + \pi_T^{(T+1)} \\ &= \sum_{t\in[T-1]} q_t^{(T)} + \pi_T^{(T)} = \Sigma_q^{(T)} = 1 \end{aligned} \tag{6.77}$$

This proves the induction step.

b) The proof is by induction in t, starting with base case $t = T$, for which the claim holds trivially. Suppose it holds for $t + 1$; its shown below this implies it holds for t:

$$\begin{aligned} \mathbb{P}(\mathsf{x}^{(T)} \geq t) &= \mathbb{P}(\mathsf{x}^{(T)} = t) + \mathbb{P}(\mathsf{x}^{(T)} \geq t + 1) \\ &= q_t^{(T)} + \pi_{t+1}^{(T)} = p_t^{(T)} \pi_t^{(T)} + \pi_{t+1}^{(T)} \\ &= p_t^{(T)} \pi_t^{(T)} + \bar{p}_t^{(T)} \pi_t^{(T)} = \pi_t^{(T)} \end{aligned} \tag{6.78}$$

c) This follows from the elementary fact that the expectation of a non-negative discrete RV may be expressed in terms of its CCDF as $\mathbb{E}[\mathsf{x}] = \sum_t \mathbb{P}(\mathsf{x} \geq t)$.

Applying these notions to the sequence $\hat{p} = \tilde{p}^{SS-S}$ with $T = t_{\tilde{p}}^{(1)}$ establishes Eq. (6.44). □

Proof of Prop. 16. This proof requires the following result: for finite (c, x, y),

$$\lim_{n\uparrow\infty} n\left[\left(1-\frac{c}{n}\right)^{x} - \left(1-\frac{c}{n}\right)^{x+y}\right] = yc. \tag{6.79}$$

In turn consider the three asserted Claims from Prop. 16.

Claim i): SS-R and SS-S. By Fact 1 and Thm. 13:

$$\frac{\tilde{p}_t^{\mathrm{SS-S}}}{p_t^{\mathrm{SS-R}}} = \frac{(\mathbb{E}[\mathsf{n}_0^{e,*}] - n_0^*)\bar{s}^t + n_0^*}{n\bar{s}^t - \frac{\bar{s}}{s}(1-\bar{s}^t)} \Bigg/ \frac{\mathbb{E}[\mathsf{n}_0^{e,*}]}{n} \tag{6.80}$$

$$= \frac{n\bar{s}^t\mathbb{E}[\mathsf{n}_0^{e,*}] - n\bar{s}^t n_0^* + nn_0^*}{n\bar{s}^t\mathbb{E}[\mathsf{n}_0^{e,*}] - \frac{\bar{s}}{s}\mathbb{E}[\mathsf{n}_0^{e,*}] + \frac{\bar{s}^{t+1}}{s}\mathbb{E}[\mathsf{n}_0^{e,*}]}. \tag{6.81}$$

Next consider the two Scenarios in turn.

Scenario i) In this case $\mathbb{E}[\mathsf{n}_0^{e,*}] = O(ns + (2ns\bar{s}\log(n))^{\frac{1}{2}})$. Substitution into Eq. (6.81) and rearrangement gives $\tilde{p}_t^{\mathrm{SS-S}}/p_t^{\mathrm{SS-R}} =$

$$\begin{aligned}\frac{\tilde{p}_t^{\mathrm{SS-S}}}{p_t^{\mathrm{SS-R}}} &= (ns\bar{s}^t + \bar{s}^t(2ns\bar{s}\log(n))^{\frac{1}{2}} - n_0^*\bar{s}^t + n_0^*) \\ &\quad\times\left[ns\bar{s}^t + \bar{s}^t(2ns\bar{s}\log(n))^{\frac{1}{2}} + (\bar{s}^{t+1} - \bar{s})\right. \\ &\quad\left.+ (2ns\bar{s}\log(n))^{\frac{1}{2}}\left(\frac{\bar{s}^{t+1}}{ns} - \frac{\bar{s}}{ns}\right)\right]^{-1}.\end{aligned} \tag{6.82}$$

Substituting $s(n) = \frac{c}{n}$ into Eq. (6.82) and rearranging gives,

$$\begin{aligned}\frac{\tilde{p}_t^{\mathrm{SS-S}}}{p_t^{\mathrm{SS-R}}} &= \left[c\left(1-\frac{c}{n}\right)^t + \left(1-\frac{c}{n}\right)^t\left(2c\left(1-\frac{c}{n}\right)\log(n)\right)^{\frac{1}{2}}\right. \\ &\quad\left.- n_0^*\left(1-\frac{c}{n}\right)^t + n_0^*\right] \\ &\quad/\left[c\left(1-\frac{c}{n}\right)^t + \left(1-\frac{c}{n}\right)^t\left(2c\left(1-\frac{c}{n}\right)\log(n)\right)^{\frac{1}{2}}\right. \\ &\quad+\left(\left(1-\frac{c}{n}\right)^{t+1} - 1 + \frac{c}{n}\right) + \left(2c\left(1-\frac{c}{n}\right)\log(n)\right)^{\frac{1}{2}} \\ &\quad\left.\times\left(\frac{\left(1-\frac{c}{n}\right)^{t+1} - \left(1-\frac{c}{n}\right)}{c}\right)\right].\end{aligned} \tag{6.83}$$

Taking the limit as $n \uparrow \infty$, using Eq. (6.79), establishes the result.

Scenario ii) In this case $\mathbb{E}[\mathsf{n}_0^{e,*}] = n_0^* + (n - n_0^*)(1 - \bar{s}^{n_0^*}) = n - \bar{s}^{n_0^*}(n - n_0^*)$. Substitution into

Eq. (6.81) and rearranging gives $\tilde{p}_t^{\mathrm{SS-S}}/p_t^{\mathrm{SS-R}} = P(n)/(Q(n)+R(n))$ where, using $s(n)=c/n$,

$$\begin{aligned}
P(n) &= n\left(\left(1-\frac{c}{n}\right)^t - \left(1-\frac{c}{n}\right)^{n_0^*+t}\right) \\
&\quad + n_0^*\left(\left(1-\frac{c}{n}\right)^{n_0^*+t} - \left(1-\frac{c}{n}\right)^t\right) + n_0^* \\
Q(n) &= n\left[\left(1-\frac{c}{n}\right)^t - \left(1-\frac{c}{n}\right)^{n_0^*+t} + \frac{1}{c}\left[\left(1-\frac{c}{n}\right)^{t+1}\right.\right. \\
&\quad \left.\left. - \left(1-\frac{c}{n}\right) + \left(1-\frac{c}{n}\right)^{n_0^*+1} - \left(1-\frac{c}{n}\right)^{t+n_0^*+1}\right]\right] \\
R(n) &= n_0^*\left[\left(1-\frac{c}{n}\right)^{t+n_0^*} + \right. \\
&\quad \left.\frac{1}{c}\left(\left(1-\frac{c}{n}\right)^{t+n_0^*+1} - \left(1-\frac{c}{n}\right)^{n_0^*+1}\right)\right].
\end{aligned}$$

Given Eq. (6.79) it follows that

$$\lim_{n\uparrow\infty} P(n) = (c+1)n_0^*,\ \lim_{n\uparrow\infty} Q(n) = cn_0^*,\ \lim_{n\uparrow\infty} R(n) = n_0^*. \tag{6.84}$$

It follows that $\lim_{n\uparrow\infty} \tilde{p}_t^{\mathrm{SS-S}}/p_t^{\mathrm{SS-R}} = \lim_{n\uparrow\infty} P(n)/(Q(n)+R(n)) = 1$. This establishes Claim *i*).

Claim ii): SS-R and SS-C. Since each SS-C sample removes a vertex from $\mathcal{V}_0\backslash\mathcal{V}_0^{e,*}$ it follows that $p_t^{\mathrm{SS-C}} = \mathbb{E}[\mathsf{n}_0^{e,*}]/(n-t+1)$, and as such, Claim *ii*) follows directly from Fact 1.

Claim iii): SS-C and SS-S. This follows immediately from Claims *i*) and *ii*). □

Chapter 7: Common greedy wiring and rewiring heuristics do not guarantee maximum assortative graphs of given degree

7.1 Introduction

7.1.1 Motivation

The assortativity of a graph (Newman [2002]) is the correlation of the degrees of the endpoints of a randomly selected edge. High degree vertices tend to be connected to high (low) degree vertices in positively (negatively) assortative graphs.

One (of many) practical implications of assortativity is in graph search, e.g., searching a (often large order) graph for (one or all) vertices of maximum (or at least large) degree Avrachenkov et al. [2014]: the work presented in Chaps. 3 and 4 has studied the performance impact of assortativity on search heuristics such as sampling and random walks. Finding such vertices in large graphs has diverse applications, including viral marketing in social networks and network robustness analysis Kempe et al. [2003], Cohen et al. [2001], among numerous others.

The motivation for this chapter is the problem of identifying a collection of graphs, all from the class of graphs with a given degree sequence, with the assortativity of the graphs in the collection varying from the minimum to the maximum possible within that class. The performance impact of the assortativity on the search heuristic may be studied by running the heuristic on all graphs in the collection. Given this objective, the first step is to identify graphs with extremal assortativity within the class. This chapter examines two greedy heuristics for finding maximum assortative graphs within a class: graph rewiring and wiring.

7.1.2 Related Work

There is an extensive literature on extremization of assortativity over different graph classes; this section briefly covers the most pertinent points of this literature, focusing on the distinctions between the work presented in this chapter and the prior work.

Assortativity. Newman [2002] introduced (graph) assortativity which is denoted $\alpha \in [-1, +1]$. Van Mieghem et al. [2010] showed perfect assortativity ($\alpha = 1$) is only possible in regular graphs, while any complete bipartite graph $K_{m,n}$ ($m \neq n$) is perfectly disassortative ($\alpha = -1$). There is a large literature on network degree correlations and assortativity (e.g., Orsini et al. [2015]), and on graphs with extremal assortativity within a class (e.g., Kincaid et al. [2016]).

Joint Degree Matrix (JDM). The generation of random graphs with a particular JDM (also called a 2K-series) has been the subject of a number of recent papers. Stanton and Pinar [2012] and Orsini et al. [2015] have proposed random edge rewiring as a method of sampling graphs with a given JDM, while Gjoka et al. [2015] has introduced a random wiring method for constructing these graphs. However, there is no means known to us by which JDMs may be efficiently enumerated, and therefore there is no easy means to maximize assortativity, which is a statistic of the JDM, short of enumerating all (in this chapter, simple and connected) graphs with a given degree sequence.

Rewiring. The meta-graph for a degree sequence, with a vertex for each connected simple graph with that degree sequence and an edge connecting graphs related by rewiring a pair of edges, was studied by Taylor [1981]; in particular, he showed this meta-graph to be connected (Thm. 3.3) extending an earlier result by Ryser [1957] for simple graphs. This fact is used in Sec. 7.2.

Following Rysler's work, rewiring heuristics for sampling graphs with a particular degree sequence (e.g., Kannan et al. [1999], Maslov and Sneppen [2002], Orsini et al. [2015]) have been introduced. Rewiring heuristics have also been proposed by Newman [2003], Xulvi-Brunet and Sokolov [2005], Van Mieghem et al. [2010], and Winterbach et al. [2012] along others for changing a graph's assortativity. The first three of these algorithms, being purely stochastic, cannot efficiently maximize assortativity. Winterbach's algorithm uses a guided rewiring technique to maximize assortativity. However, this technique does not maintain graph connectivity, as its rewirings are a subset of those explored by rewiring heuristic A (see Sec. 7.2.1), and therefore Winterbach's algorithm does not necessarily maximize assortativity.

Wiring. Li et al. [2005] introduced a greedy wiring heuristic for constructing a graph with maximum assortativity over the set of simple connected graphs with a target degree sequence. Kincaid et al. [2016] argues wiring a minimally or maximally assortative connected simple graph is NP-hard and proposes a heuristic which is shown numerically to perform near optimally in minimizing graph assortativity. Winterbach et al. [2012], Zhou et al. [2008], and Meghanathan [2016] have also proposed methods unconstrained by graph connectivity of wiring maximally assortative graphs. This chapter examines Li's heuristic further in Sec. 7.3.

Graph enumeration and generation. The results in this chapter where achieved using geng, a tool in the nauty package created by McKay and Piperno [2014], to generate all simple connected graphs of a given order.

7.1.3 Notation

Let $a \equiv b$ denote equal by definition. Let $[n]^+$ denote $\{1, \ldots, n\}$ for $n \in \mathbb{N}$. A graph of order n is denoted $G = (\mathcal{V}, \mathcal{E})$, with vertices $\mathcal{V} = [n]^+$ and edges $\mathcal{E}$; size is denoted by $m = |\mathcal{E}|$. A directed edge between vertices i and j is denoted (ij), and an undirected edge is denoted ij or $\{ij\}$.[1] Let d_i denote the degree of vertex i, $\mathbf{d} = (d_i, i \in \mathcal{V})$ denote a degree sequence, and $\mathbf{d}_G = (d_i, i \in \mathcal{V})$ the degree sequence for graph G. Additionally, let $\mathrm{Uni}(\mathcal{V})$ denote the uniform distribution over vertex set $\mathcal{V}$, $\mathrm{Var}(d_{\mathsf{w}})$ be the variance of the degree of a randomly selected vertex $\mathsf{w} \sim \mathrm{Uni}(\mathcal{V})$, and $\mathrm{Corr}(d_{\mathsf{u}}, d_{\mathsf{v}})$ be the correlation between the degrees of random vertices u and v.

The collection of distinct unlabeled undirected simple connected graphs of order $n \in \mathbb{N}$ is denoted $\mathcal{W}^{(n)}$. Let $\mathscr{D}^{(n)} \equiv \bigcup_{G \in \mathcal{W}^{(n)}} \mathbf{d}_G$ be the collection of degree sequences found in graph collection $\mathcal{W}^{(n)}$, and let $\mathcal{W}^{(n)}_{\mathbf{d}} \equiv \{G \in \mathcal{W}^{(n)} | \mathbf{d}_G = \mathbf{d}\}$ be the graphs in $\mathcal{W}^{(n)}$ with degree sequence $\mathbf{d}$, henceforth referred to as the *degree class* $\mathbf{d}$. It follows that $(\mathcal{W}^{(n)}_{\mathbf{d}}, \mathbf{d} \in \mathscr{D}^{(n)})$ is the partition of $\mathcal{W}^{(n)}$ by the degree sequence $\mathbf{d}$.

The S-metric and assortativity, for $G = (\mathcal{V}, \mathcal{E}) \in \mathcal{W}^{(n)}$, are defined below.

Definition 25. *The S-metric Li et al. [2005] is,*

$$s(G) \equiv \sum_{ij \in \mathcal{E}} d_i d_j. \tag{7.1}$$

This implies for $\{\mathsf{uv}\} \sim \mathrm{Uni}(\mathcal{E})$ an edge selected uniformly at random $\mathbb{E}[d_{\mathsf{u}} d_{\mathsf{v}}] = \frac{1}{|\mathcal{E}|} \sum_{ij \in \mathcal{E}} d_i d_j$. It follows that the assortativity Newman [2002] is, for $\mathsf{w} \sim \mathrm{Uni}(\mathcal{V})$ a vertex selected uniformly at random,

$$\alpha(G) \equiv \mathrm{Corr}(d_{\mathsf{u}}, d_{\mathsf{v}}) = \frac{s(G)/|\mathcal{E}| - \mathbb{E}[d_{\mathsf{w}}]^2}{\mathrm{Var}(d_{\mathsf{w}})}. \tag{7.2}$$

It is evident that maximizing the S-metric is equivalent to maximizing assortativity over a degree class:

$$\mathcal{W}^{(n)}_{\mathbf{d},\mathrm{opt}} \equiv \underset{G \in \mathcal{W}^{(n)}_{\mathbf{d}}}{\mathrm{argmax}} \left(s(G)\right) = \underset{G \in \mathcal{W}^{(n)}_{\mathbf{d}}}{\mathrm{argmax}} \left(\alpha(G)\right). \tag{7.3}$$

Here, $\mathcal{W}^{(n)}_{\mathbf{d},\mathrm{opt}}$ denotes those graphs achieving maximum assortativity over $\mathcal{W}^{(n)}_{\mathbf{d}}$. If there is a unique such graph it is denoted $G^{(n)}_{\mathbf{d},\mathrm{opt}}$.

[1] Except in Sec. 7.3 where Alg. 6's undirected pedges are listed as an ordered pair.

7.1.4 Contributions and outline

The rest of the chapter is organized as follows. Sec. 7.2 studies several greedy rewiring heuristics, each with the goal of identifying a graph of maximum assortativity over the degree class. Counterexamples are presented showing each of the rewiring heuristics may fail to identify such a graph. Sec. 7.3 examines the greedy wiring heuristic of Li and Alderson Li et al. [2005] designed to identify a graph of maximum assortativity over the degree class. This chapter presents a counterexample showing the heuristic may fail to produce a graph in the degree class, and also present a counterexample showing that the heuristic may produce a graph in the class that is not maximally assortative. Both Sec. 7.2 and Sec. 7.3 present tabulations of the number of counterexamples of the various types for graphs of order up to $n = 9$. Sec. 7.4 contains concluding remarks.

7.2 Rewiring

For a degree class $\mathcal{W}_{\mathbf{d}}^{(n)}$ and an initial graph $G_0 \in \mathcal{W}_{\mathbf{d}}^{(n)}$, a rewiring heuristic produces a sequence of graphs $(G_0, \ldots, G_T)$, each graph in $\mathcal{W}_{\mathbf{d}}^{(n)}$, where G_{t+1} is obtained from G_t by selecting two edges (connecting four distinct vertices) from G_t, say (ij, kl), and forming G_{t+1} with (ij, kl) replaced by either (ik, jl) or (il, jk). Any rewiring is invalid if the resulting graph is either disconnected or has multiple edges, i.e., not in $\mathcal{W}_{\mathbf{d}}^{(n)}$.

7.2.1 Greedy rewiring heuristics

A stochastic rewiring heuristic involves selecting the two edges (ij, kl) at random. While simple to implement, stochastic rewiring has no guarantee on efficiency. This observation lead to the focus of this chapter, greedy rewiring heuristics. Fix $G \in \mathcal{W}_{\mathbf{d}}^{(n)}$ and four distinct vertices $\{i, j, k, l\}$, such that G has edges (ij, kl). Rewire edges (ij, kl) to produce either graph $G' = G(ik, jl)$ or $G' = G(il, jk)$; the arguments denote the two new edges replacing edges (ij, kl). Rewiring induces a change in the S-metric:

$$\Delta_{G,G'} \equiv s(G') - s(G) = \begin{cases} (d_i d_k + d_j d_l) - (d_i d_j + d_k d_l), & G' = G(ik, jl) \\ (d_i d_l + d_j d_k) - (d_i d_j + d_k d_l), & G' = G(il, jk) \end{cases} \tag{7.4}$$

Given edges (ij, kl), $\Delta_{G,G'}$ in Eq. (7.4) is the scalar difference of the S-metric of G and one of the two possible rewirings, $G(ik, jl)$ or $G(il, jk)$, producing distinct G'. The greedy rewiring heuristics introduced below explore both rewirings.

Three greedy rewiring heuristics are developed using $\Delta_{G,G'}$; each yields a neighborhood $\mathcal{N}_{\mathbf{d},G}^{(H)}$ of

graphs in a meta-graph on $\mathcal{W}_{\mathbf{d}}^{(n)}$ (defined below), where each $G' \in \mathcal{N}_{\mathbf{d},G}^{(H)}$ is achieved by a heuristic H approved single rewiring of G.

- *A:* Improve (or maintain) $s(G)$: $\mathcal{N}_{\mathbf{d},G}^{(A)}$ holds all simple connected graphs G' obtainable by a single rewiring of G such that $\Delta_{G,G'} \geq 0$.
- *B:* Maximize Δ: $\mathcal{N}_{\mathbf{d},G}^{(B)}$ holds all simple connected graphs G' obtainable by a single rewiring of G such that $\Delta_{G,G'}$ is maximum over all G'.
- *C:* Improve and maximize: $\mathcal{N}_{\mathbf{d},G}^{(C)}$ holds all simple connected graphs G' obtainable by a single rewiring of G such that $\Delta_{G,G'} \geq 0$ and $\Delta_{G,G'}$ is maximum over all G'.

7.2.2 Meta-graphs for a degree class

Meta-graphs are graphs with vertices corresponding to the (simple and connected) non-isomorphic graphs in a degree class $\mathcal{W}_{\mathbf{d}}^{(n)}$, for a given degree sequence $\mathbf{d} \in \mathscr{D}^{(n)}$. Taylor [1981] defined the *undirected* meta-graph $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)} = (\mathcal{W}_{\mathbf{d}}^{(n)}, \hat{\mathcal{E}}_{\mathbf{d}})$, where edges are added between all pairs of graphs related by edge rewiring, i.e., $\{G, G'\} \in \hat{\mathcal{E}}_{\mathbf{d}}$ iff $G' = G(ik, jl)$ or $G' = G(il, jk)$ for some pair of edges (ij, kl). Taylor proved (Thm. 3.3) that $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)}$ is connected. Thus, any graph in $\mathcal{W}_{\mathbf{d}}^{(n)}$ is obtainable, starting from any other graph in $\mathcal{W}_{\mathbf{d}}^{(n)}$, through a sequence of rewirings, where each graph in the sequence is simple and connected. Note $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)}$ may have self-loops as rewiring G may yield G' isomorphic to G.

Rewiring heuristics A, B, and C each correspond to *directed* meta-graphs. First, label each graph $G \in \mathcal{W}_{\mathbf{d}}^{(n)}$ with its assortativity $\alpha(G)$ (alternately, $s(G)$). Next, for each heuristic $H \in \{A, B, C\}$, form the directed meta-graph $\hat{\mathcal{G}}_{\mathbf{d},H}^{(n)} \equiv (\mathcal{W}_{\mathbf{d}}^{(n)}, \hat{\mathcal{E}}_{\mathbf{d}}^{(H)})$, where $\hat{\mathcal{E}}_{\mathbf{d}}^{(H)} \equiv \{(G, G') \in \hat{\mathcal{E}}_{\mathbf{d}} | G' \in \mathcal{N}_{\mathbf{d},G}^{(H)}\}$. That is, each rewiring heuristic is represented by retaining (and orienting) the subset of edges in Taylor's meta-graph $\hat{\mathcal{G}}_{\mathbf{d}}^{(n)}$ that satisfy the heuristic.

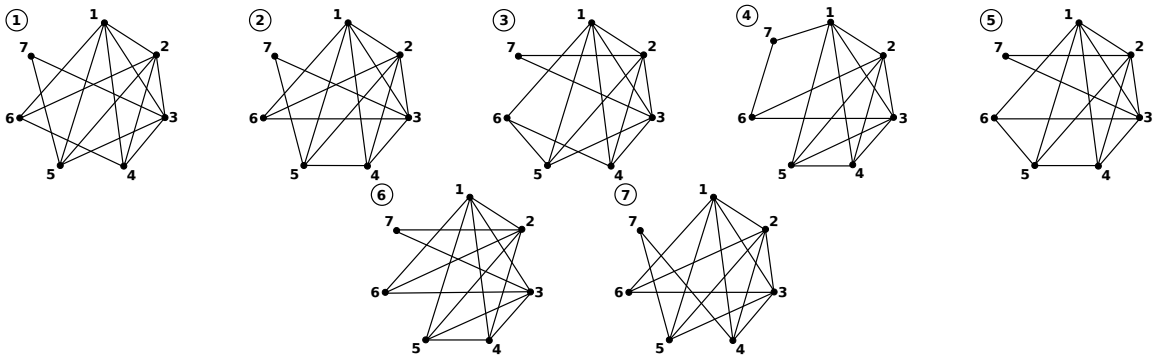

Figure 7.1: From top left to bottom right the graphs corresponding to vertices 1, 2, 3, 4, 5, 6, and 7 in $\hat{\mathcal{G}}_{\mathbf{d}}^{(7)}$, see Fig. 7.3.

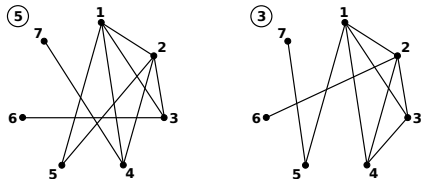

Figure 7.2: From left to right: *i*) initial graph $G_{0,B}$, node 5 in $\hat{\mathcal{G}}^{(7)}_{\mathbf{d}}$ *ii*) target graph $G^{(7)}_{\acute{\mathbf{d}},\text{opt}}$, node 3 in $\hat{\mathcal{G}}^{(7)}_{\acute{\mathbf{d}}}$, see Fig. 7.3.

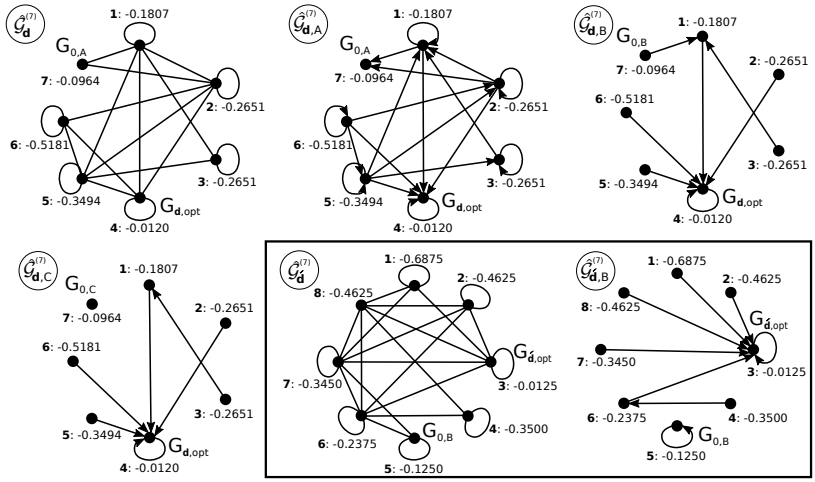

Figure 7.3: The meta-graphs above are, from top left to bottom right: *i*) $\hat{\mathcal{G}}^{(7)}_{\mathbf{d}}$, *ii*) $\hat{\mathcal{G}}^{(7)}_{\mathbf{d},A}$, *iii*) $\hat{\mathcal{G}}^{(7)}_{\mathbf{d},B}$, *iv*) $\hat{\mathcal{G}}^{(7)}_{\mathbf{d},C}$ with $\mathbf{d} = (5,5,5,4,4,3,2)$, *v*) $\hat{\mathcal{G}}^{(7)}_{\acute{\mathbf{d}}}$, *vi*) $\hat{\mathcal{G}}^{(7)}_{\acute{\mathbf{d}},B}$ with $\acute{\mathbf{d}} = (4,4,3,3,2,1,1)$. The number to the right of each vertex id is the assortativity of the corresponding graph.

7.2.3 Rewiring heuristic counterexamples

One might hope that (one or more of) the rewiring heuristics would provide a guarantee that, for any initial graph $G_0 \in \mathcal{W}^{(n)}_{\mathbf{d}}$, there exists a directed path, following the heuristic, from G_0 to one or more graphs in $\mathcal{W}^{(n)}_{\mathbf{d},\text{opt}}$. Unfortunately, all three heuristics can fail to achieve this goal, as shown by the counterexamples below. A counterexample for heuristic $H \in \{A, B, C\}$ identifies a $(n, \mathbf{d}, G_0)$ triple, with $n \in \mathbb{N}$, $\mathbf{d} \in \mathscr{D}^{(n)}$, and $G_0 \in \mathcal{W}^{(n)}_{\mathbf{d}}$, such that there is no path from G_0 to any graph in $\mathcal{W}^{(n)}_{\mathbf{d},\text{opt}}$ in the directed meta-graph $\hat{\mathcal{G}}^{(n)}_{\mathbf{d},H}$.

Note that these heuristics do not specify a particular rewiring, i.e., each heuristic identifies, in general, a collection of possible neighborhood graphs $\mathcal{N}^{(H)}_G$, where each graph in the neighborhood is consistent with the heuristic. Thus, a counterexample for the heuristic has the property that the heuristic H would fail to achieve the target set for *any* possible choice of $G' \in \mathcal{N}^{(H)}_G$, for each G "reachable" from G_0.

Counterexample 1. *Fix order* $n = 7$*, degree sequence* $\mathbf{d} = (5,5,5,4,4,3,2)$*, and initial graph*

$G_{0,A} \in \mathcal{V}_{\mathbf{d}}^{(7)}$ (graph 7 in Fig. 7.1). The (unique) graph with maximum assortativity, $G_{\mathbf{d},\text{opt}}^{(7)}$, is graph 4 in Fig. 7.1. The meta-graph $\hat{\mathcal{G}}_{\mathbf{d}}^{(7)}$ and directed meta-graph under heuristic A, $\hat{\mathcal{G}}_{\mathbf{d},A}^{(7)}$ are graph 1 and 2 in Fig. 7.3. There is no path from $G_{0,A}$ to $G_{\mathbf{d},\text{opt}}^{(7)}$ in $\hat{\mathcal{G}}_{\mathbf{d},A}^{(7)}$, and hence no path in $\hat{\mathcal{G}}_{\mathbf{d},C}^{(7)}$ (graph 4 in Fig. 7.3). Thus, $(n, \mathbf{d}, G_{0,A})$ is a counterexample for heuristics A and C.

This counterexample asserts that graph $G_{0,A}$ has locally (i.e., over graphs adjacent to $G_{0,A}$ in $\hat{\mathcal{G}}_{\mathbf{d},A}^{(7)}$) maximal but not globally (i.e., over $\mathcal{W}_{\mathbf{d}}^{(7)}$) maximum assortativity. To see that $G_{0,A}$ is locally maximal, Tab. 7.1 lists possible pairs of edges from $G_{0,A}$ which if rewired as $G' = G_{0,A}(ik, jl)$ or $G' = G_{0,A}(il, jk)$ maintain graph simplicity and connectivity: $\Delta_{G_{0,A},G'} < 0$ for each possible G'.

Table 7.1: Rewirings of edge pairs (ij, kl) (left) of $G_{0,A}$, along with $\Delta_{G_{0,A},G'}$ for $G' = G_{0,A}(ik, jl)$ (middle) or $G' = G_{0,A}(il, jk)$ (right). Bold entries maximize $\Delta_{G_{0,A},G'}$, * indicates rewirings which violate graph simplicity or connectivity.

(ij, kl)	(ik, jl)	$\Delta_{G_{0,A},G'}$	(il, jk)	$\Delta_{G_{0,A},G'}$
$(43, 57)$	$(45, 37)$	-2	$(47, 35)$	*
$(42, 57)$	$(45, 27)$	-2	$(47, 25)$	*
$(41, 57)$	$(45, 17)$	-2	$(47, 15)$	*
$(47, 53)$	$(45, 73)$	-2	$(43, 75)$	*
$(47, 52)$	$(45, 72)$	-2	$(42, 75)$	*
$(47, 51)$	$(45, 71)$	-2	$(41, 75)$	*
$(47, 63)$	$(46, 73)$	$\mathbf{-1}$	$(43, 67)$	*
$(47, 62)$	$(46, 72)$	$\mathbf{-1}$	$(42, 76)$	*
$(47, 61)$	$(46, 71)$	$\mathbf{-1}$	$(41, 76)$	*
$(57, 63)$	$(56, 73)$	$\mathbf{-1}$	$(53, 67)$	*
$(57, 62)$	$(56, 72)$	$\mathbf{-1}$	$(52, 76)$	*
$(57, 61)$	$(56, 71)$	$\mathbf{-1}$	$(51, 76)$	*

Counterexample 2. *Fix order $n = 7$, degree sequence $\acute{\mathbf{d}} = (4, 4, 3, 3, 2, 1, 1)$, and initial graph $G_{0,B} \in \mathcal{W}_{\acute{\mathbf{d}}}^{(7)}$ (graph 1 in Fig. 7.2). The (unique) graph with maximum assortativity, $G_{\acute{\mathbf{d}},\text{opt}}^{(7)}$, is graph 2 in Fig. 7.2. The meta-graph $\hat{\mathcal{G}}_{\acute{\mathbf{d}}}^{(7)}$ and directed meta-graph under heuristic B, $\hat{\mathcal{G}}_{\acute{\mathbf{d}},B}^{(7)}$, are graph 5 and 6 in Fig. 7.3. There is no path from $G_{0,B}$ to $G_{\acute{\mathbf{d}},\text{opt}}^{(7)}$ in $\hat{\mathcal{G}}_{\acute{\mathbf{d}},B}^{(7)}$. Thus, $(n, \acute{\mathbf{d}}, G_{0,B})$ is a counterexample for heuristic B.*

This counterexample asserts that graph $G_{0,B}$ has locally maximal but not globally maximum assortativity. This can be seen by enumerating all possible pairs of edges from $G_{0,B}$ which if rewired maintain graph simplicity and connectivity, and showing the (unique) optimal choice to maximize $\Delta_{G,G'}$ produces a new graph G' isomorphic to $G_{0,B}$; this enumeration is omitted due to space

constraints. The isomorphism between $G_{0,B}$ and G' produces the self-loop in $\hat{\mathcal{G}}^{(7)}_{\mathbf{d},B}$ at the vertex corresponding to $G_{0,B}$.

C. Exp. 1 and C. Exp. 2 were found via exhaustive search. The class of non-isomorphic simple connected graphs of n vertices and degree sequence $\mathbf{d}$, $\mathcal{W}^{(n)}_{\mathbf{d}}$ were enumerated using the tool `geng` McKay and Piperno [2014]. Letting G_i denote the i_{th} graph in $\mathcal{W}^{(n)}_{\mathbf{d}}$ corresponding to the i_{th} vertex in meta-graph $\hat{\mathcal{G}}^{(n)}_{\mathbf{d},H}$, under greedy rewiring heuristic H, all non-isomorphic edge rewirings of G_i are enumerated. For each rewiring G'_i of G_i that satisfies heuristic H one checks for isomorphism of G'_i with $G_k \in \mathcal{W}^{(n)}_{\mathbf{d}}$. If G'_i is isomorphic with G_k, a directed edge (i,k) is added to meta-graph $\hat{\mathcal{G}}^{(n)}_{\mathbf{d},H}$. Upon completing this procedure for all $G \in \mathcal{W}^{(n)}_{\mathbf{d}}$, one checks that a path exists from each vertex in $\hat{\mathcal{G}}^{(n)}_{\mathbf{d},H}$ to a vertex in $\mathcal{W}^{(n)}_{\mathbf{d},\text{opt}}$. This procedure was used to generate Tab. 7.2, which counts the number of counterexamples for each of the greedy rewiring heuristics.

Table 7.2: Rewiring heuristic counterexample counts: The number of distinct graphs $|\mathcal{W}^{(n)}|$ and degree sequences $|\mathscr{D}^{(n)}|$, followed by the number of distinct graphs ($\#G$) and degree sequences ($\#\mathbf{d}$) that are counterexamples for heuristics A, B, C, for $n \in \{6,7,8,9\}$.

	overall		heuristic A		heuristic B		heuristic C	
n	$\lvert\mathcal{W}^{(n)}\rvert$	$\lvert\mathscr{D}^{(n)}\rvert$	$\#G$	$\#\mathbf{d}$	$\#G$	$\#\mathbf{d}$	$\#G$	$\#\mathbf{d}$
6	112	68	0	0	0	0	0	0
7	853	236	2	2	1	1	2	2
8	11,117	863	13	12	15	8	20	12
9	261,080	3,137	149	80	1045	67	1100	80

7.3 Wiring

If a degree sequence $\mathbf{d}$ satisfies the Erdős Gallai theorem there exists one or more simple connected graphs with that degree sequence, i.e., $\mathcal{W}^{(n)}_{\mathbf{d}} \neq \emptyset$ Erdős and Gallai [1960]. Given such a $\mathbf{d}$, a wiring heuristic produces a sequence of graphs $(\tilde{G}_0, \ldots, \tilde{G}_T)$, with $\tilde{G}_0$ the empty graph, such that $\tilde{G}_{t+1}$ is formed from $\tilde{G}_t$ by adding one edge, subject to the constraint that no vertex $j \in \mathcal{V}$ is ever assigned a degree exceeding its target d_j. It is typical to consider each vertex j in graph $\tilde{G}_t$ as having d_j "stubs" of which some number $\tilde{d}_j$ hold edges, and the remainder, $\delta_j \equiv d_j - \tilde{d}_j$, are available for wiring. The goal of the wiring heuristics is to obtain a graph of maximum assortativity, i.e., $\tilde{G}_T \in \mathcal{W}^{(n)}_{\mathbf{d},\text{opt}}$ given $\mathbf{d}$.

7.3.1 Greedy wiring heuristic

Li et al. [2005] developed the elegant greedy wiring heuristic in Alg. 6 which, given a degree sequence $\mathbf{d}$ is intended to produce a graph $\tilde{G}$ that is *i*) feasible, i.e., that is in $\mathcal{W}^{(n)}_{\mathbf{d}}$, and *ii*) optimal, i.e.,

in $\mathcal{W}^{(n)}_{\mathbf{d},\mathrm{opt}}$. Although the heuristic performs well on most inputs, the following section will present counterexamples demonstrating neither property is guaranteed for all $\mathbf{d}$.

Each potential edge, hereafter a "pedge", is denoted by the ordered pair (ij) with $i < j$. The basic idea is to select from set of all pedges $\mathcal{O}$ those with the largest endpoint degree product, $\mathcal{M}$ (Alg. 6), after removing from $\mathcal{O}$ and $\mathcal{M}$ pedges in $\mathcal{M}$ without available (unwired) stubs $\mathcal{F}$ (Alg. 6). If pedges remain then further ties are broken by first (then second) selecting the pedge (ij) with the most unwired stubs δ_i (δ_j). Vertices $[n]^+$ are partitioned into $\mathcal{R}, \mathcal{Q}$, where $\mathcal{R}$ ($\mathcal{Q}$) holds any vertex with one or more (no) edges. If the pedge (ij) has $i \in \mathcal{R}$ and $j \in \mathcal{Q}$ then the edge is added and vertex j moves from $\mathcal{Q}$ to $\mathcal{R}$ (Alg. 6). Else $\mathcal{R}$ holds both i and j and one must check the "tree condition", $(d_{\mathcal{Q}} \neq (2|\mathcal{Q}| - \delta_{\mathcal{R}}))$, and "disconnected cluster condition", $(\delta_{\mathcal{R}} \neq 2)$, in Alg. 6.

The "tree condition" is required since at any point in wiring the graph, connecting the vertices in $\mathcal{Q}$ to the $\delta_{\mathcal{R}} := \sum_{k\in\mathcal{R}} \delta_k$ free stubs in $\mathcal{R}$ requires $\delta_{\mathcal{R}}$ acyclic graphs and at least $2|\mathcal{Q}| - \delta_{\mathcal{R}}$ free stubs in $\mathcal{Q}$. As $\delta_{\mathcal{R}}$ decreases, the number of free stubs in $\mathcal{Q}$ required to connect the vertices in $\mathcal{Q}$ to the free stubs in $\mathcal{R}$ increases. Letting $\delta_{\mathcal{Q}} := \sum_{k\in\mathcal{Q}} \delta_k$, if an edge is added between vertices i and j in $\mathcal{R}$ which results in $d_{\mathcal{Q}} < 2|\mathcal{Q}| - \delta_{\mathcal{R}}$ then there are not enough free stubs in $\mathcal{Q}$ to connect all the vertices in $\mathcal{Q}$ to those in $\mathcal{R}$, entailing that $\tilde{G}$ is disconnected. The "disconnected cluster condition" is required since wiring an edge between the only two free stubs in $\mathcal{R}$ entails $\tilde{G}$ is disconnected, as no additional vertices in $\mathcal{Q}$ can be attached to those in $\mathcal{R}$ (see Li et al. [2005]). Regardless of whether or not the conditions in Alg. 6 are satisfied, the pedge (ij) being considered for wiring is removed from the set $\mathcal{O}$ in Alg. 6.

Algorithm 6 Greedy wiring heuristic (adapted from Li et al. [2005])

require: $\mathbf{d} = (d_1, \ldots, d_n)$ with $d_1 \geq \cdots \geq d_n$
$\mathcal{R} := \{1\}$, $\mathcal{Q} := \{2, \ldots, n\}$, $\tilde{\mathcal{E}} := \{\}$, $\tilde{G} := (\mathcal{R}, \tilde{\mathcal{E}})$, $\mathcal{O} := \{(ij) : 1 \leq i < j \leq n\}$
while $\mathcal{O} \neq \emptyset$ **do**
    $\mathcal{M} := \operatorname{argmax}_{(ij)\in\mathcal{O}}(d_i d_j)$
    $\mathcal{F} := \{(ij) \in \mathcal{M} : \delta_i \delta_j = 0\}$
    $\mathcal{O} := \mathcal{O} \backslash \mathcal{F}$, $\mathcal{M} := \mathcal{M} \backslash \mathcal{F}$
    if $\mathcal{M} \neq \emptyset$ **then**
        $\mathcal{M}' := \operatorname{argmax}_{(ij)\in\mathcal{M}} \delta_i$
        Select $(ij) \in \operatorname{argmax}_{(ij)\in\mathcal{M}'} \delta_j$
        if $i \in \mathcal{R}$ and $j \in \mathcal{Q}$ **then**
            $\tilde{\mathcal{E}} := \tilde{\mathcal{E}} \cup \{ij\}$, $\mathcal{R} := \mathcal{R} \cup \{j\}$, $\mathcal{Q} := \mathcal{Q} \backslash \{j\}$
        else
            $d_{\mathcal{Q}} := \sum_{k\in\mathcal{Q}} d_k$, $\delta_{\mathcal{R}} := \sum_{k\in\mathcal{R}} \delta_k$
            if $(d_{\mathcal{Q}} \neq (2|\mathcal{Q}| - \delta_{\mathcal{R}})) \wedge (\delta_{\mathcal{R}} \neq 2)$ **then**
                $\tilde{\mathcal{E}} := \tilde{\mathcal{E}} \cup \{ij\}$
            end if
        end if
        $\mathcal{O} := \mathcal{O} \backslash \{(ij)\}$
    end if
end while
return $\tilde{G}$

Alg. 6 is underspecified in Alg. 6, i.e., there may be multiple edges after sorting $\mathcal{O}$ by $d_i d_j$, δ_i, and δ_j, and no guidance is provided in Li et al. [2005] for selecting a pedge in such a case. To compensate for this, the implementation of Alg. 6 in this chapter selects *all* possible pedge choices, via a breadth first search, returning all possible graphs $\tilde{G}$ that may result from a valid pedge selection in Alg. 6. A degree sequence $\mathbf{d}$ is considered to be a counterexample for *i*) feasibility if none of the returned graphs are in $\mathcal{W}_{\mathbf{d}}^{(n)}$, and *ii*) optimality if at least one returned graph is in $\mathcal{W}_{\mathbf{d}}^{(n)}$, yet none are in $\mathcal{W}_{\mathbf{d},\text{opt}}^{(n)}$.

7.3.2 Wiring heuristic counterexamples

Counterexample 3. *Fix* $n = 6$ *and* $\mathbf{d} = (5, 4, 4, 4, 4, 3)$ *(which satisfies the Erdős Gallai theorem). The graph* $\tilde{G}$ *returned by Alg. 6 does not have the target degree sequence, i.e.,* $\mathbf{d}_{\tilde{G}} \neq \mathbf{d}$*, and thus* $\tilde{G}$ *is not feasible, i.e.,* $\tilde{G} \notin \mathcal{W}_{\mathbf{d}}^{(6)}$.

Proof. Tab. 7.3 gives the sequence of wirings satisfying $(ij) \in \operatorname{argmax}_{(ij)\in\mathcal{O}}(d_i d_j)$, illustrated in Fig. 7.4. The first four edges added, namely $(12), \ldots, (15)$, have identical priority as $d_i d_j$, δ_i, and δ_j are equal for each. These four may be added in any order without affecting the resulting graph. The next edges added will be (23), (24), (25), (34), (35), (45). Finally, (16) will be added, leaving the only two free stubs in the graph on vertex 6, which can only be wired via a self-loop, thereby violating the requirement that $\tilde{G}$ be simple. □

Table 7.3: Subset of edge wirings for C. Exp. 3. The first set of rows correspond to wirings which are optimal at wiring step 1. The second set of rows are optimal wirings at wiring step 5. The final row is the only legal at wiring at step 11.

(ij)	$d_i d_j$	δ_i	δ_j
(12)	20	5	4
(13)	20	5	4
(14)	20	5	4
(15)	20	5	4
(23)	16	3	3
(24)	16	3	3
(25)	16	3	3
(34)	16	3	3
(35)	16	3	3
(45)	16	3	3
(16)	15	1	3

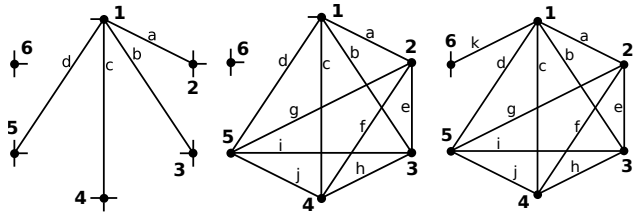

Figure 7.4: Snapshots of the graph wiring in C. Exp. 3 for $n = 6$ and $\mathbf{d} = (5, 4, 4, 4, 4, 3)$ where the edges are added in alphabetical order: From left to right *i)* $\tilde{G}_4$, *ii)* $\tilde{G}_{10}$, *iii)* $\tilde{G}_{11}$.

Counterexample 4. *Fix* $n = 8$ *and* $\mathbf{d} = (6, 4, 4, 4, 4, 3, 2, 1)$ *(which satisfies the Erdős Gallai theorem). The graph* $\tilde{G}$ *returned by Alg. 6 is feasible, but its assortativity is not maximum and thus* $\tilde{G}$ *is not optimal, i.e.,* $\tilde{G} \notin \mathcal{W}^{(8)}_{\mathbf{d},\text{opt}}$.

Proof. The proof is similar to C. Exp. 3. The partially wired graphs at steps 5, 11, 12, and 14 are shown in Fig. 7.5. The returned graph $\tilde{G} = \tilde{G}_{14}$ achieves the target degree sequence $\mathbf{d}$, however its assortativity is not optimal. Namely, $\alpha(\tilde{G}_{14}) = -0.04886$ while $\alpha(G^{(8)}_{\mathbf{d},\text{opt}}) = -0.00326$, where $G^{(8)}_{\mathbf{d},\text{opt}}$ is graph 5 in Fig. 7.5. □

C. Exp. 3 and C. Exp. 4 were found by enumerating all degree sequences of order n which satisfy the Erdős Gallai theorem. Given a degree sequence $\mathbf{d}$ of n vertices, Alg. 6 is used to wire target degree sequence $\mathbf{d}$. Given the breadth first search, Alg. 6 returns a set of graphs $\tilde{\mathcal{G}}$. If $\mathbf{d}_{\tilde{G}} \neq \mathbf{d}$ for all $\tilde{G} \in \tilde{\mathcal{G}}$, $\mathbf{d}$ is counted as a feasibility counter example. Otherwise, we check if there exists any $\tilde{G} \in \tilde{\mathcal{G}}$ (obeying $\mathbf{d}_{\tilde{G}} = \mathbf{d}$) for which $\alpha(\tilde{G}) = \alpha(G^{(n)}_{\mathbf{d},\text{opt}})$. If not, $\mathbf{d}$ is counted as an optimality counterexample. This procedure is used to generate the counts of feasibility and optimality counterexamples in Tab. 7.4.

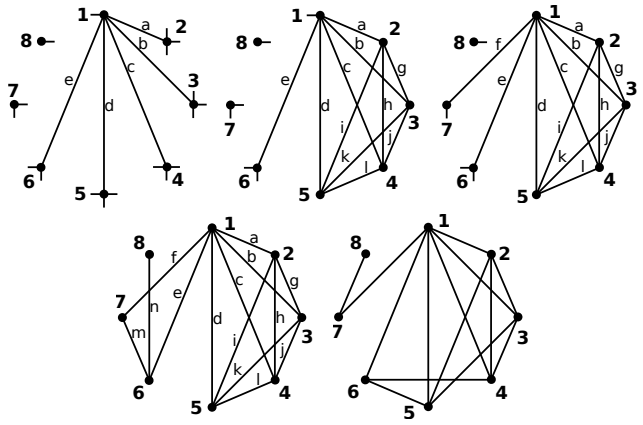

Figure 7.5: Snapshots of the graph wiring in C. Exp. 4 for $n = 8$ and $\mathbf{d} = (6, 4, 4, 4, 4, 3, 2, 1)$ where the edges are added in alphabetical order: From top left to bottom right *i)* $\tilde{G}_5$, *ii)* $\tilde{G}_{11}$, *iii)* $\tilde{G}_{12}$, *iv)* $\tilde{G}_{14}$, and *v)* the maximally assortative graph $G^{(8)}_{\mathbf{d},\mathrm{opt}}$.

Table 7.4: Wiring heuristic counterexample counts: The number of degree sequences $|\mathscr{D}^{(n)}|$, the number of degree sequences for which the returned graph is not feasible, and (if feasible) is not optimal, for $n \in \{5, 6, 7, 8, 9\}$.

n	$\lvert\mathscr{D}^{(n)}\rvert$	feasibility	optimality
5	19	0	0
6	68	2	0
7	236	16	0
8	863	91	4
9	3,137	443	36

7.4 Conclusion

The main point of this chapter is to demonstrate the failure of natural greedy heuristics, for both graph rewiring and wiring, to produce connected simple graphs with maximum assortativity over an arbitrary target degree class. Many open questions remain, such as how the relative prevalence of the various classes of counterexamples scale with n. One possible direction for future work is to seek to characterize common structural properties of the degree sequences $\mathbf{d} \in \mathscr{D}^{(n)}$ comprising the four types of counterexamples given above.

Chapter 8: Online estimation for finding a near-maximum value in a large list of numerical data

8.1 Introduction

8.1.1 Near-maximum values in large unsorted datasets

This chapter addresses a problem of interest in mining a large unsorted list of numerical data: given only the knowledge of the dataset's *distribution family* and *order* (i.e., its number of elements) find an element of near-maximum value. For *small* order datasets this is a trivial sorting problem. It is non-trivial, however, for *large* order datasets where such enumeration is not possible. Natural examples include extreme values in large datasets of, say, macroeconomic indicators, genomic data and bioassays, geospatial data, astronomical data, and the large graphs typically studied by network science and big data.

Although several motivating examples are drawn from graph contexts, the problem formulation and algorithms are not graph-specific, as the approach in this chapter does not make explicit use of the graph structure. Still, the problem formulation is pertinent to large graphs in contexts where querying the graph for a specific node returns only the degree, effectively reducing the graph to a large unsorted list of numerical data.

8.1.2 Dataset query and required sample size

While large unsorted datasets cannot be efficiently enumerated, they may often be *queried*. This chapter defines a dataset to be *subject to query* if: *i*) it provides a unique index to each entry in the dataset, *ii*) the set of indices is known, and *iii*) the value of the entry x_i is discovered by supplying index i to the query engine. Thus, $x_i = \mathsf{query}_X(i)$ is a query of the value x_i of the entry with index $i \in [n]$ in data set X.

This chapter assumes a unit cost is incurred by each query of data set X, and the only available knowledge of the dataset is the set of indices, $[n]$, or its order, $n \equiv |X|$. Given these assumptions, and a choice of the "nearness parameter" $\delta \in [0, 1)$, this chapter quantifies the number of queries required to find a value in $I(\delta) \equiv \{i \in [n] | x_i / x_{\max} \geq 1 - \delta\}$, the set of indices of elements near $x_{\max}$, under random sampling without replacement (RSW). As it stands this problem is ill-posed as it presumes knowledge of the maximum value in the dataset, $x_{\max}$. This chapter shows that *online estimation* can be used to avoid this presumption.

8.1.3 Contributions and outline

The contribution of this chapter is offline and online estimators for the number of queries required to find an element in $I(\delta)$, and a preliminary evaluation of their performance. In particular, it is shown that these estimators give reasonably accurate estimates of the required sample size for synthetic binomial and Zipf/zeta distributed datasets modeling the degrees of Erdős-Rényi (ER) and Barabasi-Albert (BA) graphs. In addition, this chapter observes the impact of the dataset order (n) and the dataset's distribution on the required sample size. Specifically, *i*) for a given "nearness" value δ, the sample size is observed to decrease in the order of the dataset, *ii*) light-tailed dataset distributions require far smaller sample sizes than heavy-tailed dataset distributions. Intuitively, the small number of very large values for heavy-tails makes it harder to find near-maximum values relative to the difficulty for light-tails.

This chapter is organized as follows. Results for general (continuous) RVs are given in Sec. 8.2; results for specific distributions are in Sec. 8.3. The performance metrics and sampling algorithms introduced in this chapter are in Sec. 8.4, with results for synthetic datasets where either the distribution parameters are known or only the distribution family is known *a priori*. Related work is given in Sec. 8.5, and Sec. 8.6 concludes the chapter.

8.2 General Continuous Distributions

This section gives a Taylor series approximation of the expectation of the ratio of a pair of order statistics.

8.2.1 Notation

Write $a \equiv b$ to denote a and b are equal by definition. For $n \in \mathbb{N}$ let $[n] \equiv \{1, \ldots, n\}$. Sans-serif capital letters, e.g., X denote random variables (RV). The cumulative distribution function (CDF) and probability density function (PDF) for X are denoted $F_{\mathsf{X}}(x)$ and $f_{\mathsf{X}}(x)$. Expectation, variance, and covariance are denoted $\mu_{\mathsf{X}} \equiv \mathbb{E}[\mathsf{X}]$, $\sigma^2_{\mathsf{X}} \equiv \mathrm{Var}(\mathsf{X})$, and $\gamma_{\mathsf{X}_1,\mathsf{X}_2} \equiv \mathrm{Cov}(\mathsf{X}_1, \mathsf{X}_2)$.

Let $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ be a collection of n IID RVs. Then $(\mathsf{X}_{1:n}, \ldots, \mathsf{X}_{n:n})$ denotes a permutation of this collection into non-decreasing order, $\mathsf{X}_{1:n} \leq \cdots \leq \mathsf{X}_{n:n}$, and $\mathsf{X}_{i:n}$ is the i^{th} smallest variable. Fix integer k with $1 \leq k < n$. Let $(\mathsf{X}_1, \ldots, \mathsf{X}_k)$ be the truncation of the collection to the first k values. Then $\mathsf{X}_{k:k} \equiv \max(\mathsf{X}_1, \ldots, \mathsf{X}_k)$ and $\mathsf{X}_{n:n} \equiv (\mathsf{X}_1, \ldots, \mathsf{X}_n)$ denote the maximum value over the first k entires and the entire set respectively. Write $\mu_{\mathsf{X}}(k) \equiv \mathbb{E}[\mathsf{X}_{k:k}]$, $\sigma^2_{\mathsf{X}}(k) \equiv \mathrm{Var}(\mathsf{X}_{k:k})$, and $\gamma_{\mathsf{X}}(k, n) \equiv \mathrm{Cov}(\mathsf{X}_{k:k}, \mathsf{X}_{n:n})$.

8.2.2 Taylor series approximations of the expectation of a ratio

Let $(\mathsf{X}_1, \mathsf{X}_2)$ be a pair of continuous RVs, with $\mathsf{X}_2 \neq 0$ almost surely, means $(\mu_{\mathsf{X}_1}, \mu_{\mathsf{X}_2})$, variances $(\sigma^2_{\mathsf{X}_1}, \sigma^2_{\mathsf{X}_2})$, and covariance $\gamma_{\mathsf{X}_1,\mathsf{X}_2}$, and consider the expectation of the ratio, i.e., $\mathbb{E}\left[\frac{\mathsf{X}_1}{\mathsf{X}_2}\right]$. The following result is found in Van Kempen and Van Vliet [2000] and Seltman [2017].

Proposition 21 (Van Kempen and Van Vliet [2000], Seltman [2017])**.** *The first and second order Taylor series approximations of* $\mathbb{E}\left[\frac{\mathsf{X}_1}{\mathsf{X}_2}\right]$ *around* $(\mu_{\mathsf{X}_1}, \mu_{\mathsf{X}_2})$ *are*

$$\mathbb{E}\left[\frac{\mathsf{X}_1}{\mathsf{X}_2}\right] = \frac{\mu_{\mathsf{X}_1}}{\mu_{\mathsf{X}_2}} + b_1 \tag{8.1}$$

$$\mathbb{E}\left[\frac{\mathsf{X}_1}{\mathsf{X}_2}\right] = \frac{\mu_{\mathsf{X}_1}}{\mu_{\mathsf{X}_2}} - \frac{\gamma_{\mathsf{X}_1,\mathsf{X}_2}}{\mu^2_{\mathsf{X}_2}} + \frac{\sigma^2_{\mathsf{X}_2}\mu_{\mathsf{X}_1}}{\mu^3_{\mathsf{X}_2}} + b_2 \tag{8.2}$$

where (b_1, b_2) *are "small" remainder terms.*

8.2.3 Covariance bounds for order statistics

The following two results give lower and upper bounds on the covariance $\gamma_{\mathsf{X}}(k, n)$ between $\mathsf{X}_{k:k}$ and $\mathsf{X}_{n:n}$ $(1 \leq k < n)$.

Proposition 22. *The covariance of* $\mathsf{X}_{k:k}, \mathsf{X}_{n:n}$ *is bounded by*

$$0 \leq \gamma_{\mathsf{X}}(k, n) \leq \sigma_{\mathsf{X}}(k)\sigma_{\mathsf{X}}(n). \tag{8.3}$$

Proof. The lower bound holds as both $\mathsf{X}_{k:k}$ and $\mathsf{X}_{n:n}$ are nondecreasing functions of the underlying collection of RVs $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$. The upper bound holds by the upper bound on the correlation, i.e., $\frac{\gamma_{\mathsf{X}}(k,n)}{\sigma_{\mathsf{X}}(k)\sigma_{\mathsf{X}}(n)} \leq 1$, which in turn follows from the Cauchy-Schwarz inequality. □

While the above bounds are quite general, the next result leverages specific properties of the two order statistics.

Proposition 23. *The covariance of* $\mathsf{X}_{k:k}, \mathsf{X}_{n:n}$ *is bounded by*

$$\begin{aligned} \gamma_{\mathsf{X}}(k, n) &\geq \sigma^2_{\mathsf{X}}(k) - \mu_{\mathsf{X}}(k)(\mu_{\mathsf{X}}(n) - \mu_{\mathsf{X}}(k)), \\ \gamma_{\mathsf{X}}(k, n) &\leq \sigma^2_{\mathsf{X}}(n) + \mu_{\mathsf{X}}(n)(\mu_{\mathsf{X}}(n) - \mu_{\mathsf{X}}(k)). \end{aligned} \tag{8.4}$$

Proof. Let $\nu_{\mathsf{X}}(k,n) \equiv \mathbb{E}[\mathsf{X}_{k:k}\mathsf{X}_{n:n}]$. Then

$$\begin{aligned}\gamma_{\mathsf{X}}(k,n) &\equiv \mathrm{Cov}(\mathsf{X}_{k:k},\mathsf{X}_{n:n}) \\ &\equiv \mathbb{E}[\mathsf{X}_{k:k}\mathsf{X}_{n:n}] - \mathbb{E}[\mathsf{X}_{k:k}]\mathbb{E}[\mathsf{X}_{n:n}] \\ &\equiv \nu_{\mathsf{X}}(k,n) - \mu_{\mathsf{X}}(k)\mu_{\mathsf{X}}(n). \end{aligned} \tag{8.5}$$

Lower bound: observe

$$\nu_{\mathsf{X}}(k,n) = \mathbb{E}[\mathsf{X}_{k:k}\max\{\mathsf{X}_{k:k},\tilde{\mathsf{X}}_{n-k:n-k}\}], \tag{8.6}$$

where $\tilde{\mathsf{X}}_{n-k:n-k}$ is the maximum of a separate IID sequence of RVs $(\tilde{\mathsf{X}}_1,\ldots,\tilde{\mathsf{X}}_{n-k})$. As $\max\{\cdot,\cdot\}$ is lower bounded by each of its arguments:

$$\begin{aligned}\nu_{\mathsf{X}}(k,n) &\geq \max\{\mathbb{E}[\mathsf{X}_{k:k}\mathsf{X}_{k:k}],\mathbb{E}[\mathsf{X}_{k:k}\tilde{\mathsf{X}}_{n-k:n-k}]\} \\ &= \max\{\mathbb{E}[\mathsf{X}_{k:k}^2],\mathbb{E}[\mathsf{X}_{k:k}]\mathbb{E}[\tilde{\mathsf{X}}_{n-k:n-k}]\} \\ &= \max\{\sigma_{\mathsf{X}}^2(k)+\mu_{\mathsf{X}}(k)^2,\mu_{\mathsf{X}}(k)\mu_{\mathsf{X}}(n-k)\}. \end{aligned} \tag{8.7}$$

This in turn yields the covariance lower bounds

$$\begin{aligned}\gamma_{\mathsf{X}}(k,n) \geq \max\{&\sigma_{\mathsf{X}}^2(k)-\mu_{\mathsf{X}}(k)(\mu_{\mathsf{X}}(n)-\mu_{\mathsf{X}}(k)), \\ &\mu_{\mathsf{X}}(k)(\mu_{\mathsf{X}}(n-k)-\mu_{\mathsf{X}}(n))\}. \end{aligned} \tag{8.8}$$

The second quantity in the above expression, however, is negative, and so is inferior to the lower bound $\gamma_{\mathsf{X}}(k,n) \geq 0$. *Upper bound:* $\mathbb{E}[\mathsf{X}_{k:k}\mathsf{X}_{n:n}] \leq \mathbb{E}[\mathsf{X}_{n:n}\mathsf{X}_{n:n}] = \sigma_{\mathsf{X}}^2(n)+\mu_{\mathsf{X}}(n)^2$. □

8.2.4 Taylor series approx. of expected ratio of order statistics

The first key result follows from the above results.

Proposition 24. *Let $(\mathsf{X}_1,\ldots,\mathsf{X}_n)$ be IID. The first and second order Taylor series approximations of the expected ratio of order statistics $\mathsf{X}_{k:k},\mathsf{X}_{n:n}$, i.e., $\mathbb{E}\left[\frac{\mathsf{X}_{k:k}}{\mathsf{X}_{n:n}}\right]$, around $(\mu_{\mathsf{X}}(k),\mu_{\mathsf{X}}(n))$ are*

$$\begin{aligned}\mathbb{E}\left[\frac{\mathsf{X}_{k:k}}{\mathsf{X}_{n:n}}\right] &= \frac{\mu_{\mathsf{X}}(k)}{\mu_{\mathsf{X}}(n)} + c_1 \\ \mathbb{E}\left[\frac{\mathsf{X}_{k:k}}{\mathsf{X}_{n:n}}\right] &= \frac{\mu_{\mathsf{X}}(k)}{\mu_{\mathsf{X}}(n)} - \frac{\gamma_{\mathsf{X}}(k,n)}{\mu_{\mathsf{X}}(n)^2} + \frac{\sigma_{\mathsf{X}}^2(n)\mu_{\mathsf{X}}(k)}{\mu_{\mathsf{X}}(n)^3} + c_2, \end{aligned} \tag{8.9}$$

where (c_1, c_2) *are "small" remainder terms, and* $\gamma_{\mathsf{X}}(k,n)$ *has the following upper and lower bounds:*

$$\begin{aligned} &\leq \min\{\sigma^2_{\mathsf{X}}(n) + \mu_{\mathsf{X}}(n)(\mu_{\mathsf{X}}(n) - \mu_{\mathsf{X}}(k)), \sigma_{\mathsf{X}}(k)\sigma_{\mathsf{X}}(n)\} \\ &\geq \max\{\sigma^2_{\mathsf{X}}(k) - \mu_{\mathsf{X}}(k)(\mu_{\mathsf{X}}(n) - \mu_{\mathsf{X}}(k)), 0\}. \end{aligned} \tag{8.10}$$

8.3 Specific continuous distributions

We apply the general results from Sec. 8.2 to the case of the normal and Pareto distributions in Sec. 8.3.1 and Sec. 8.3.2, respectively.

8.3.1 Normal RVs

Let $(\mathsf{Z}_1, \ldots, \mathsf{Z}_n)$ be a collection of n IID standard normal RVs, where $\mathsf{Z}_{n:n}$ denotes the maximum value, $\mu_{\mathsf{Z}}(n) \equiv \mathbb{E}[\mathsf{Z}_{n:n}]$, and $\sigma^2_{\mathsf{Z}}(n) \equiv \mathrm{Var}(\mathsf{Z}_{n:n})$. Similarly, let $(\mathsf{W}_1, \ldots, \mathsf{W}_n)$ be a collection of n IID normal RVs with parameters (μ_0, σ_0), where $\mathsf{W}_{n:n}$ denotes the maximum value, $\mu_{\mathsf{W}}(n) \equiv \mathbb{E}[\mathsf{W}_{n:n}]$, and $\sigma^2_{\mathsf{W}}(n) \equiv \mathrm{Var}(\mathsf{W}_{n:n})$. The following result is in Arnold et al. [2008].

Theorem 15 (Arnold et al. [2008])**.** *Let* $(\mathsf{Z}_1, \ldots, \mathsf{Z}_n)$*,* $\mathsf{Z}_{n:n}$*,* $\mu_{\mathsf{Z}}(n)$*, and* $\sigma^2_{\mathsf{Z}}(n)$ *be as above. Then* $\mu_{\mathsf{Z}}(n) \approx \tilde{\mu}_{\mathsf{Z}}(n)$ *and* $\sigma^2_{\mathsf{Z}}(n) \approx \tilde{\sigma}^2_{\mathsf{Z}}(n)$*, where*

$$\tilde{\mu}_{\mathsf{Z}}(n) \equiv \sqrt{2\log n} - \frac{\log(4\pi \log n)}{2\sqrt{2\log n}} + \frac{\gamma}{\sqrt{2\log n}}, \tag{8.11}$$

$$\tilde{\sigma}^2_{\mathsf{Z}}(n) \equiv \frac{\pi^2}{12\log n}, \tag{8.12}$$

and where $\gamma \approx 0.577$ *is the Euler-Mascheroni constant.*

The approximation accuracy is shown in Fig. 8.1 (top), which shows Monte-Carlo simulations ($m = 10^4$ samples per point) of the maximum of n IID standard normal RVs vs. n, i.e., each point is the sample mean (left) and sample standard deviation (right) of the m maximums of random n-vectors.

The following is an immediate corollary of Thm. 15.

Corollary 5. *Let* $(\mathsf{W}_1, \ldots, \mathsf{W}_n)$*,* $\mathsf{W}_{n:n}$*,* $\mu_{\mathsf{W}}(n)$*, and* $\sigma^2_{\mathsf{W}}(n)$ *be as above. Then* $\mu_{\mathsf{W}}(n) \approx \tilde{\mu}_{\mathsf{W}}(n) \equiv \mu_0 + \sigma_0\tilde{\mu}_{\mathsf{Z}}(n)$ *and* $\sigma^2_{\mathsf{W}}(n) \approx \tilde{\sigma}^2_{\mathsf{W}}(n) \equiv \sigma_0^2\tilde{\sigma}^2_{\mathsf{Z}}(n)$*.*

Proof. Observe that $\mathsf{W}_i = \mu_0 + \sigma_0\mathsf{Z}_i$ for $i \in [n]$, and as such

$$\begin{aligned} \mathbb{E}[\mathsf{W}_{n:n}] &= \mathbb{E}[\max_{i\in[n]}(\mu_0 + \sigma_0\mathsf{Z}_i)] = \mathbb{E}[\mu_0 + \sigma_0\mathsf{Z}_{n:n}] \\ \mathrm{Var}(\mathsf{W}_{n:n}) &= \mathrm{Var}(\max_{i\in[n]}(\mu_0 + \sigma_0\mathsf{Z}_i)) = \mathrm{Var}(\mu_0 + \sigma_0\mathsf{Z}_{n:n}). \end{aligned} \tag{8.13}$$

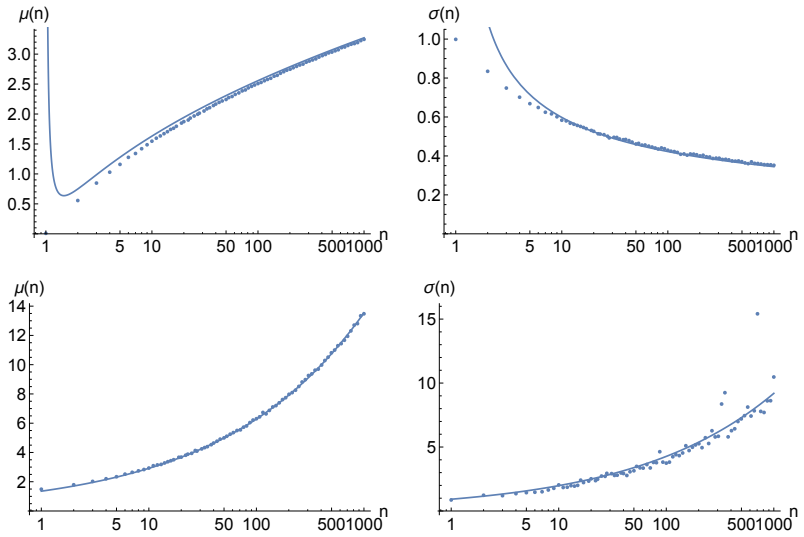

Figure 8.1: *Top left:* Monte-Carlo simulation results for $\mu_{\mathsf{Z}}(n)$ (points) and its approximation $\tilde{\mu}_{\mathsf{Z}}(n)$ (solid line) in Thm. 15 of the maximum value of n IID standard normal RVs. *Top right:* same, but for $\sigma_{\mathsf{Z}}(n)$ and its approximation $\tilde{\sigma}_{\mathsf{Z}}(n)$ in Thm. 15. *Bottom:* same as top but for the Pareto distribution with parameters $y_0 = 1$ and $\alpha = 3$, and approximations $\tilde{\mu}_{\mathsf{Y}}(n), \tilde{\sigma}^2_{\mathsf{Y}}(n)$ in Thm. 16.

□

Prop. 24 and Cor. 5 immediately yield Cor. 6.

Corollary 6. *Let* $(\mathsf{W}_1, \ldots, \mathsf{W}_n)$ *be IID normal RVs. The first and second order Taylor series approximations of the expected ratio of order statistics* $\mathsf{W}_{k:k}, \mathsf{W}_{n:n}$*, i.e.,* $\mathbb{E}\left[\frac{\mathsf{W}_{k:k}}{\mathsf{W}_{n:n}}\right]$*, around* $(\tilde{\mu}_{\mathsf{W}}(k), \tilde{\mu}_{\mathsf{W}}(n))$ *given by Prop. 24 are in turn approximated using the expressions given in Cor. 5, i.e., with* $\tilde{\mu}_{\mathsf{W}}(m)$ *and* $\tilde{\sigma}^2_{\mathsf{W}}(m)$ *replacing* $\mu_{\mathsf{X}}(m)$ *and* $\sigma^2_{\mathsf{X}}(m)$*, respectively.*

The accuracy of the approximations for the normal distribution from Thm. 15 is shown in Fig. 8.2 (top). *i*) the second-order approx. using the lower bound $\gamma(k, n) \geq 0$ is highly accurate, but exceeds the upper bound of one for $k/n \approx 1$, *ii*) the second order approx. using the upper and lower bounds in Prop. 23 are far less accurate than the bounds in Prop. 22, although the lower bound from Prop. 23 outperforms the lower bound from Prop. 22 for $k/n \approx 1$, *iii*) the first-order approx. $\mu(k)/\mu(n)$ is quite accurate relative to the second-order approximations, due to the inaccuracy of the covariance bounds.

8.3.2 Pareto RVs

Let $(\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$ be a collection of n IID Pareto RVs, each with CDF $F_{\mathsf{Y}}(y) = 1 - \left(\frac{y_0}{y}\right)^{\alpha}$, support $y \geq y_0$, and parameters $y_0 > 0$ and $\alpha > 2$ (assumed to ensure finite variance). Recall $\mathsf{Y}_{n:n}$ denotes

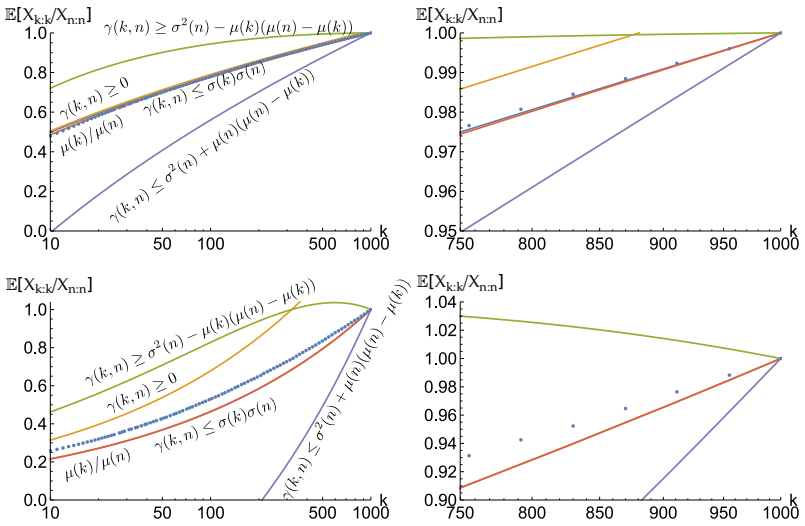

Figure 8.2: Left: the expected ratio of the order statistics for standard normal RVs $\mathbb{E}\left[\frac{\mathsf{W}_{k:k}}{\mathsf{W}_{n:n}}\right]$ (top) and for Pareto RVs $\mathbb{E}\left[\frac{\mathsf{Y}_{k:k}}{\mathsf{Y}_{n:n}}\right]$ (bottom) vs. k for $n = 1000$. Monte-Carlo simulations (averaged over $m = 10^4$ realizations) are shown as points, and the approximations from Cor. 6 (normal) and Cor. 7 (Pareto) are shown as lines. Right: details from the top right corner of the left side plots.

the maximum value, $\mu_{\mathsf{Y}}(n) \equiv \mathbb{E}[\mathsf{Y}_{n:n}]$, and $\sigma^2_{\mathsf{Y}}(n) \equiv \mathrm{Var}(\mathsf{Y}_{n:n})$. The following well-known theorem relies on EVT and is included for completeness; see, Wolpert [2013], §1.3.

Theorem 16. *Let* $(\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$, $\mathsf{Y}_{n:n}$, $\mu_{\mathsf{Y}}(n)$, *and* $\sigma^2_{\mathsf{Y}}(n)$ *be as above. Then* $\mu_{\mathsf{Y}}(n) \approx \tilde{\mu}_{\mathsf{Y}}(n)$ *and* $\sigma^2_{\mathsf{Y}}(n) \approx \tilde{\sigma}^2_{\mathsf{Y}}(n)$, *where*

$$\tilde{\mu}_{\mathsf{Y}}(n) \equiv y_0 \Gamma(1 - 1/\alpha) n^{\frac{1}{\alpha}}, \tag{8.14}$$

$$\tilde{\sigma}^2_{\mathsf{Y}}(n) \equiv y_0^2 (\Gamma(1 - 2/\alpha) - \Gamma(1 - 1/\alpha)^2) n^{\frac{2}{\alpha}}, \tag{8.15}$$

and where $\Gamma(\cdot)$ *is the gamma function.*

Proof. Let (a_n) and (b_n) be sequences, with $b_n > 0$. If $a_n + b_n y \geq y_0$ then

$$\mathbb{P}\left(\frac{\mathsf{Y}_{n:n} - a_n}{b_n} \leq y\right) = \left(1 - \left(\frac{y_0}{a_n + b_n y}\right)^{\alpha}\right)^n. \tag{8.16}$$

Set $a_n = 0$ and $b_n = y_0 n^{\frac{1}{\alpha}}$, substitute, and observe

$$\mathbb{P}\left(\frac{\mathsf{Y}_{n:n} - 0}{y_0 n^{\frac{1}{\alpha}}} \leq y\right) = \left(1 - \left(\frac{y_0}{0 + y_0 n^{\frac{1}{\alpha}} y}\right)^{\alpha}\right)^n \to \mathrm{e}^{-y^{-\alpha}}. \tag{8.17}$$

Thus, the normalized sequence of RVs $\frac{\mathsf{Y}_{n:n}-a_n}{b_n}$ converges in distribution to the Frechét RV Y_∞, with distribution $F_{\mathsf{Y}_\infty}(y) = \mathrm{e}^{-y^{-\alpha}}$, mean $\mathbb{E}[\mathsf{Y}_\infty] = \Gamma(1-1/\alpha)$ (for $\alpha > 1$) and variance $\mathrm{Var}(\mathsf{Y}_\infty) = \Gamma(1-2/\alpha) - \Gamma(1-1/\alpha)^2$ (for $\alpha > 2$). Thus,

$$\mathbb{E}\left[\frac{\mathsf{Y}_{n:n}-0}{y_0 n^{\frac{1}{\alpha}}}\right] \approx \mathbb{E}[\mathsf{Y}_\infty], \quad \mathrm{Var}\left(\frac{\mathsf{Y}_{n:n}-0}{y_0 n^{\frac{1}{\alpha}}}\right) \approx \mathrm{Var}(\mathsf{Y}_\infty), \tag{8.18}$$

and the expressions in Thm. 16 follow immediately. □

The accuracy of these approximations is shown in Fig. 8.1 (bottom). Prop. 24 and Thm. 16 immediately yield Cor. 7.

Corollary 7. *Let $(\mathsf{Y}_1, \ldots, \mathsf{Y}_n)$ be IID Pareto RVs. The first and second order Taylor series approximations of the expected ratio of order statistics $\mathsf{Y}_{k:k}, \mathsf{Y}_{n:n}$, i.e., $\mathbb{E}\left[\frac{\mathsf{Y}_{k:k}}{\mathsf{Y}_{n:n}}\right]$, around $(\tilde{\mu}_\mathsf{Y}(k), \tilde{\mu}_\mathsf{Y}(n))$ given by Prop. 24 are in turn approximated using the expressions given in Thm. 16, i.e., with $\tilde{\mu}_\mathsf{Y}(m)$ and $\tilde{\sigma}^2_\mathsf{Y}(m)$ replacing $\mu_\mathsf{X}(m)$ and $\sigma^2_\mathsf{X}(m)$, respectively.*

The accuracy of the approximations for the Pareto distribution is shown in Fig. 8.2 (bottom). Most comments on the normal case apply here. Notable distinctions for the Pareto case are: *i*) the error between the Monte-Carlo approximations and the best Taylor series approximation is significantly larger for the Pareto case than the normal, and *ii*) the first-order approximation notably outperforms all second-order approximations, but even the first-order approximation has notable error.

8.4 Sampling algorithms to find near-max. values

8.4.1 Performance metrics

Following notation from Sec. 8.1, let $X = (x_1, \ldots, x_n)$ be a dataset of known order n, let $\mathcal{X} \equiv \bigcup_{i\in[n]} x_i$ be the set of values found in X, and let $x_{\max} \equiv \max(\mathcal{X})$. The probability mass function (PMF) of X is denoted $p_X \equiv (p_X(x), x \in \mathcal{X})$, where $p_X(x) \equiv \#\{i \in [n] | x_i = x\}/n$ is the fraction of indices with value x. Thus, $p_X(x) = \mathbb{P}(x_\mathsf{U} = x)$ is the probability an index U, selected uniformly at random from $[n]$, has value x.

Recall $I(\delta) \equiv \{i \in [n] | x_i/x_{\max} \geq 1-\delta\}$ is the target index set, for $\delta \in [0, 1)$. The dataset is repeatedly queried for the value of an index, via RSW, stopping if the index of the maximum value from the sample is in the (unknown) set $I(\delta)$. Let $(\mathsf{X}_1, \ldots, \mathsf{X}_n)$ be the values of X in the random permutation by which the indices are queried. This set is extended out to length n even if sampling

stops after $k < n$ queries. After k queries the ratio of maximum value obtained over the maximum value possible is the RV $\mathsf{X}_{k:k}/\mathsf{X}_{n:n}$.

Two caveats are in order. *i*) The values $(\mathsf{X}_{k_1}, \mathsf{X}_{k_2})$ of any two samples k_1, k_2 may not be independent, but the analysis in this chapter ignores this potential dependence. *ii*) Although the dataset distribution p_X is discrete, this chapter approximates it using continuous normal and Pareto distributions. The motivation here is 1) for large n, discrete distributions may often be well-approximated by continuous distributions, and 2) EVT is easier to apply for continuous distributions.

Given a choice of the parameter δ and knowledge of both the dataset order n and the dataset distribution p_X, define

$$\begin{aligned} k(\delta) &= k(\delta, n, p_X) \equiv \min \left\{ k \in [n] \middle| \mathbb{E}\left[\frac{\mathsf{X}_{k:k}}{\mathsf{X}_{n:n}}\right] \geq 1 - \delta \right\} \\ \tilde{k}(\delta) &= \tilde{k}(\delta, n, p_X) \equiv \min \left\{ k \in [n] \middle| \frac{\tilde{\mu}(k)}{\tilde{\mu}(n)} \geq 1 - \delta \right\}. \end{aligned} \tag{8.19}$$

Here, $k(\delta)$ is the required sample size such that the expected value ratio hits the target $1 - \delta$, and $\tilde{k}(\delta)$ is an approximation of $k(\delta)$ using $\tilde{\mu}(k) \approx \mu(k) \equiv \mathbb{E}[\mathsf{X}_{k:k}]$, the approximate expected value of $\mathsf{X}_{k:k}$. Note, exact computation of $k(\delta)$ is difficult for most distributions p_X, but may be easily approximated by simulation. Define $\kappa(\delta) \equiv k(\delta)/n$ and $\tilde{\kappa}(\delta) \equiv \tilde{k}(\delta)/n$ as the *fraction* and approximate fraction of indices to be queried.

Two performance metrics for sample size k are given below:

$$r(k, n, p_X) \equiv \mathbb{E}\left[\frac{\mathsf{X}_{k:k}}{\mathsf{X}_{n:n}}\right], \; \tilde{r}(k, n, p_X) \equiv \frac{\tilde{\mu}(k)}{\tilde{\mu}(n)}. \tag{8.20}$$

Let $r(k) = r(k, n, p_X)$ and $\tilde{r}(k) = \tilde{r}(k, n, p_X)$ be the expected value ratio and approximate expected value ratio for a sample size k. Next define: $r(k(\delta))$, $r(\tilde{k}(\delta))$, $\tilde{r}(k(\delta))$, $\tilde{r}(\tilde{k}(\delta))$. In words: $r(k(\delta))$ measures the expected value ratio using the required sample size, $r(\tilde{k}(\delta))$ measures the expected value ratio using the approximate sample size, $\tilde{r}(k(\delta))$ measures the approximate expected value ratio using the required sample size, and $\tilde{r}(\tilde{k}(\delta))$ measures the approximate expected value ratio using the approximate sample size. Ideally all four of these quantities will be near to $1 - \delta$. In particular, $r(k(\delta))$ and $\tilde{r}(\tilde{k}(\delta))$ should be near $1 - \delta$ by construction. Yet $r(\tilde{k}(\delta))$ is the key metric, measuring the performance under the "ground-truth" metric $r(k)$ using the sample size approximation $\tilde{k}(\delta)$.

8.4.2 The binomial and Zipf (zeta) dataset distributions

This section introduces two dataset distributions: the binomial and Zipf/zeta. Although the problem definition and the proposed algorithm are not graph specific, graph theory is used to motivate these distributions. The binomial distribution results from the degrees of an Erdős-Rényi graph, where each edge is placed uniformly and independently at random. The Zipf/zeta distribution is a discrete power-law distribution which result from the degrees of a Barabasi-Albert graph (exhibiting preferential attachment). The normal distribution provides a good continuous approximation of the (discrete) binomial distribution (by the de Moivre–Laplace Theorem).

Corollary 8. *Let* $(\mathsf{B}_1, \ldots, \mathsf{B}_n)$ *be IID binomial RVs with parameters* (n, s)*. Let* $\mathsf{B}_{n:n}$ *denote the maximum value, with mean* $\mu_{\mathsf{B}}(n)$ *and variance* $\sigma^2_{\mathsf{B}}(n)$*. Then* $\mu_{\mathsf{B}}(n) \approx \mu_{\mathsf{W}}(n) \approx \tilde{\mu}_{\mathsf{W}}(n)$ *and* $\sigma^2_{\mathsf{B}}(n) \approx \sigma^2_{\mathsf{W}}(n) \approx \tilde{\sigma}^2_{\mathsf{W}}(n)$ *for* $\mathsf{W} \sim \mathcal{N}(\mu_0, \sigma_0)$ *in Cor. 5, with* $\mu_0 = ns$ *and* $\sigma_0 = \sqrt{ns(1-s)}$*.*

Similarly, the Pareto distribution provides a good continuous approximation of the discrete Zipf/zeta distribution.

Corollary 9. *Let* $(\mathsf{A}_1, \ldots, \mathsf{A}_n)$ *be IID Zipf/zeta RVs with power-law exponent* e*, and let* $\mathsf{A}_{n:n}$ *denote the maximum value, with mean* $\mu_{\mathsf{A}}(n)$ *and variance* $\sigma^2_{\mathsf{A}}(n)$*. Then* $\mu_{\mathsf{A}}(n) \approx \mu_{\mathsf{Y}}(n) \approx \tilde{\mu}_{\mathsf{Y}}(n)$ *and* $\sigma^2_{\mathsf{A}}(n) \approx \sigma^2_{\mathsf{Y}}(n) \approx \tilde{\sigma}^2_{\mathsf{Y}}(n)$ *for* Y *the Pareto distribution in Thm. 16, with* $y_0 = 1$ *and* $\alpha = e - 1$*.*

8.4.3 Known distribution family and parameters

This subsection assumes the dataset distribution p_X is known, and as such stopping criteria $k(\delta), \tilde{k}(\delta)$ in Sec. 8.4.1 may be computed *a priori*. The sampling algorithm is shown in Alg. 7. Alg. 7 and the performance measures, Sec. 8.4.1 and Eq. (8.20), are applied to binomial and Zipf/zeta dateset distributions; the results are in Fig. 8.3 and Fig. 8.4. In each case $m = 1000$ datasets were generated. The quantity $k(\delta)$ in Sec. 8.4.1 was estimated via Monte-Carlo simulation by averaging the values $(\mathsf{K}_1(\delta), \ldots, \mathsf{K}_m(\delta))$, where $\mathsf{K}_j(\delta) \equiv \min\{k | x_{j,k}/x_{j,\max} \geq (1-\delta)\}$ for $j \in [m]$ is the sample size required to find an element in set $I(\delta)$.

Results for Alg. 7 for binomial dataset distribution

Fig. 8.3 shows results for the binomial dataset distribution with parameters $n \in \{1000, 5000\}$ (order) and $s = 1/250$. Note $\tilde{k}(\delta)$ was computed via $\tilde{\mu}_{\mathsf{W}}(n)$ in Cor. 5 using (μ_0, σ_0) from $\mathsf{W} \sim \mathcal{N}(\mu_0, \sigma_0)$ approximation to the $\mathsf{B} \sim \mathrm{Bin}(n, s)$ in Cor. 8. *i*) There is good correspondence between $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$, meaning Cor. 5 gives a reliable estimator. *ii*) Approximations $r(\tilde{k}(\delta))$ and $\tilde{r}(k(\delta))$ are less accurate than $r(k(\delta))$ and $\tilde{r}(\tilde{k}(\delta))$. *iii*) rapidly $\kappa(\delta)$ falls off as a function of δ, especially for order

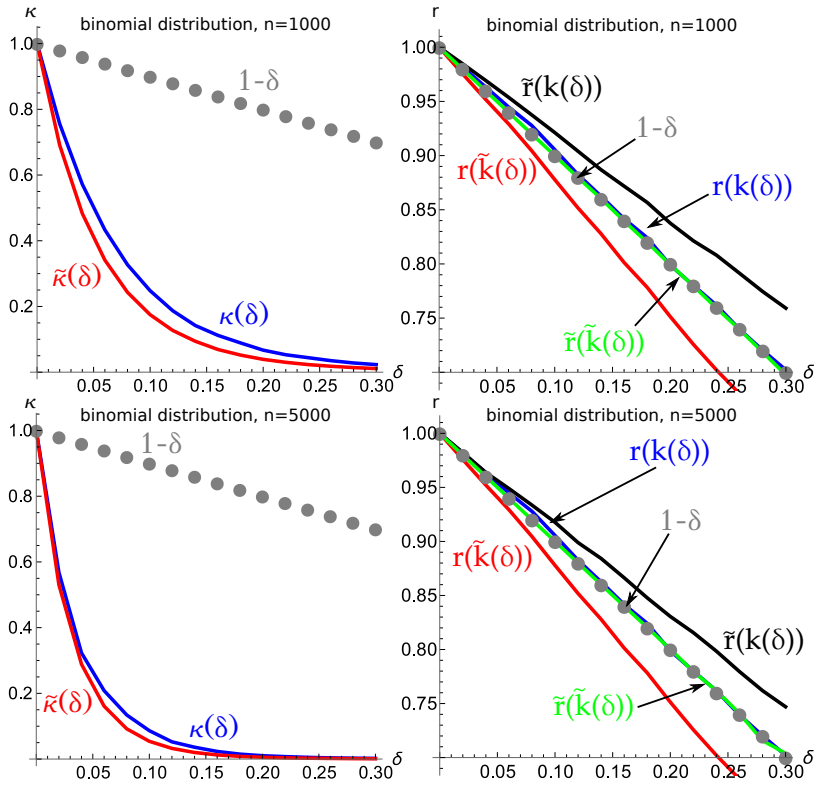

Figure 8.3: Results for Sec. 8.4.3: Alg. 7 (known parameters) for the binomial dataset distribution with $s = 1/250$ and $n = 1000$ (*top*) and $n = 5000$ (*bottom*). *Left:* fraction of indices to be sampled, $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$ in Sec. 8.4.1, vs. δ. *Right:* accuracy of $r(k(\delta))$, $r(\tilde{k}(\delta))$, $\tilde{r}(k(\delta))$, and $\tilde{r}(\tilde{k}(\delta))$ vs. δ.

$n = 5000$. For example, the bottom-left plot shows that the $\mathsf{X}_{k:k}$ obtained from sampling 10% of indices of a dataset will be on average equal to 90% of $\mathsf{X}_{n:n}$ for the dataset as a whole, whereas the required sample size for order 1000 is 20% of the indices. The decrease in the relative sample size, say $\kappa(\delta, n)$ viewed as a function of order n, is a consequence of the expected maximum value growing at order $O(\sqrt{\log n})$, c.f., Thm. 15.

Algorithm 7 Dataset sampling: known dist. family and params

1: **require** $\delta \in [0, 1)$, $n \in \mathbb{N}$, p_X
2: **initialize** $k = k(\delta)$ or $k = \tilde{k}(\delta)$ in Sec. 8.4.1
3: **initialize** random ordering $(\pi_1, \ldots, \pi_n)$
4: **query** the dataset k times with indices $(\pi_1, \ldots, \pi_k)$, returning values: $x_1 = \mathsf{query}_X(\pi_1), \ldots, x_k = \mathsf{query}_X(\pi_k)$
5: **return** max sample value $x_{\max}(k) \equiv \max(x_1, \ldots, x_k)$

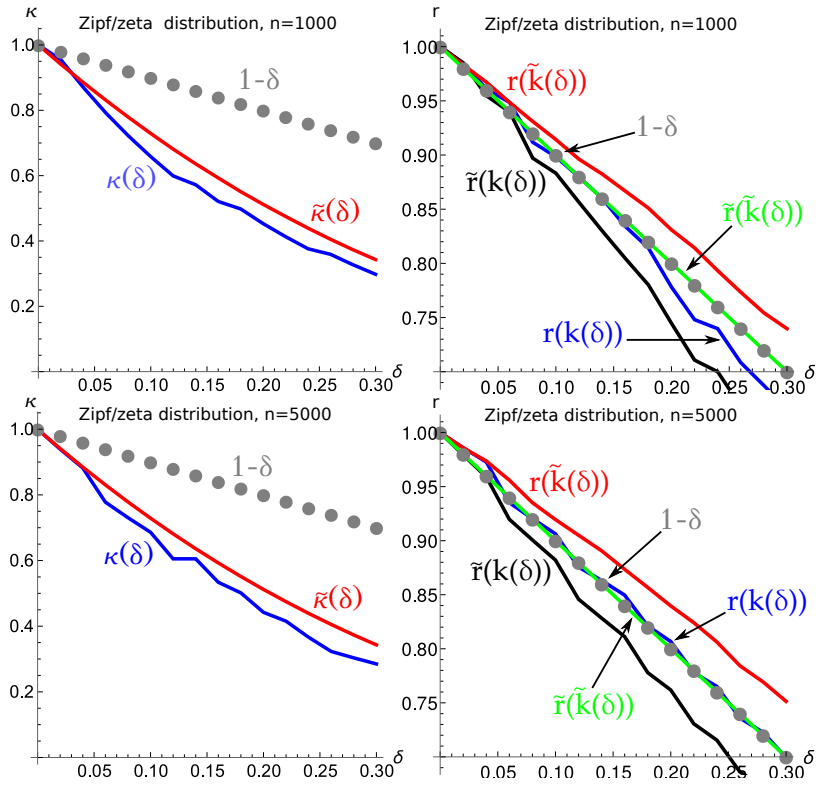

Figure 8.4: Results for Sec. 8.4.3: Alg. 7 (known parameters) Zipf/zeta with exponent $e = 4$ (Pareto disbn. exponent $\alpha = 3$), $n = 1000$ (*top*), $n = 5000$ (*bottom*). *Left:* fraction of indices to be sampled, $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$ in Sec. 8.4.1, vs. δ. *Right:* accuracy of $r(k(\delta))$, $r(\tilde{k}(\delta))$, $\tilde{r}(k(\delta))$, and $\tilde{r}(\tilde{k}(\delta))$ in vs. δ.

Results for Alg. 7 for Zipf/zeta dataset distribution

Fig. 8.4 shows results for Zipf/zeta dataset distribution with order $n \in \{1000, 5000\}$ and exponent $e = 4$, and thus the approximate Pareto distribution has exponent $\alpha = e - 1 = 3$. The quantity $\tilde{k}(\delta)$ was computed using the function $\tilde{\mu}_Y(n)$ in Thm. 16 with parameters $y_0 = 1$ and $\alpha = 3$ via Cor. 9.

The Zepf/zeta results show a good correspondence between $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$. Observe the slower rate of decay of $\kappa(\delta)$ as a function of δ. For example, results for order $n = 5000$ show that approximately 70% of the indices must be sampled in order to find a sample index with value within 90% of $\mathsf{X}_{n:n}$. This is a consequence of power-law distributions having few very large values, resulting in large (relative to the binomial case) sample sizes being required to find near-maximum values. Once again the approximations $r(\tilde{k}(\delta))$ and $\tilde{r}(k(\delta))$ are less accurate than $r(k(\delta))$ and $\tilde{r}(\tilde{k}(\delta))$.

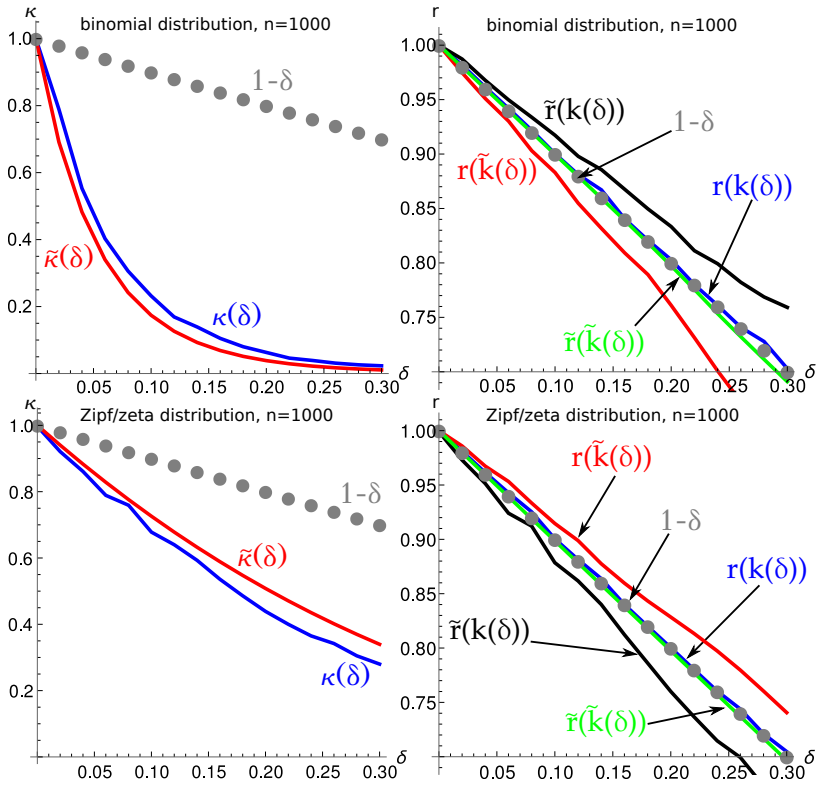

Figure 8.5: Results for Alg. 8 (unknown parameters): for binomial dataset distribution with $s = 1/250$, (*top*), Zipf/zeta dataset distribution with exponent $e = 4$ (*bottom*), each with $n = 1000$, $k_{min} = 20$, $m = 200$ datasets. *Left:* fraction of indices to be sampled, $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$ in Sec. 8.4.1, vs. δ. *Right:* accuracy of $r(k(\delta))$, $r(\tilde{k}(\delta))$, $\tilde{r}(k(\delta))$, and $\tilde{r}(\tilde{k}(\delta))$ vs. δ.

8.4.4 Known distribution family and unknown parameters

This section assumes the dataset distribution family is known but its parameters are not. These parameters are estimated online as sampling proceeds. Updated parameter values are used to revise the stopping criterion. Alg. 8 gives the sampling algorithm for the binomial dataset distribution. The idea is: *i*) estimate the mean and variance of the dataset distribution online as $(\hat{\mu}_0, \hat{\sigma}_0^2)$, *ii*) use $(\hat{\mu}_0, \hat{\sigma}_0^2)$ in the normal distribution approx. of the expected value of the maximum of k IID normal RVs, denoted $\tilde{\mu}_{\mathsf{W}}(k)$, from Cor. 5, *iii*) estimate the required sample size $\tilde{k}(\delta)$ in Sec. 8.4.1. The estimators are:

$$\hat{\mu}_0 \equiv \frac{1}{k} \sum_{i \in [k]} x_i, \quad \hat{\sigma}_0^2 \equiv \frac{1}{k-1} \sum_{i \in [k]} (x_i - \hat{\mu}_0)^2. \tag{8.21}$$

Select $k_{\min}$ in Alg. 8 to ensure noise in the estimates $(\hat{\mu}_0, \hat{\sigma}_0^2)$ does not lead to a premature decision to stop sampling.

The corresponding algorithm for the Zipf/zeta dataset distribution is similar, except the pa-

Algorithm 8 Dataset sampling: binomial w/ unknown params

require $\delta \in [0,1)$, $n \in \mathbb{N}$, $k_{\min} \in [n]$
initialize rand. ordering $(\pi_1, \ldots, \pi_n)$, $\hat{k} := k_{\min}$, $k := 1$
while $k < \hat{k}$ **do**
  $x_k = \mathsf{query}_X(\pi_k)$
  Compute $(\hat{\mu}_0, \hat{\sigma}_0)$ from Eq. (8.21)
  Compute $\tilde{k}(\delta)$ in Sec. 8.4.1 using $\tilde{\mu}_{\mathsf{W}}(k, \hat{\mu}_0, \hat{\sigma}_0)$ in Cor. 5
  $\hat{k} := \max\{k_{\min}, \tilde{k}(\delta)\}$
  $k := k + 1$
end while
return $x_{\max}(k) \equiv \max(x_1, \ldots, x_k)$

rameters of the Pareto distribution: minimum value y_0 and exponent α are estimated as $\hat{y}_0 = \min(x_1, \ldots, x_k)$, $\hat{\alpha} = \hat{e} - 1$. This assumes $x_i \neq c$ for $i \in [k]$ and $c \in \mathbb{R}$, and $\hat{e}$ is the estimated Zipf/zeta distribution parameter e solving $\frac{\zeta'(e)}{\zeta(e)} = \frac{-1}{k} \sum_{i=1}^{k} \log(x_i)$ Goldstein et al. [2004], where $\zeta'(\cdot)$ is the derivative of the Riemann zeta function. Estimators $(\hat{y}_0, \hat{\alpha})$ are used to compute $\tilde{\mu}_Y(k)$ of $\mathbb{E}[\mathsf{Y}_{k:k}]$ in Thm. 16, which in turn are used to compute $\tilde{k}(\delta)$ in Sec. 8.4.1.

Results for Alg. 8 are shown in Fig. 8.5, for the binomial dataset distribution (top) and Zipf/zeta dataset distribution (bottom). Note three points. *i*) The left-side plots of $\kappa(\delta)$ and $\tilde{\kappa}(\delta)$ from Sec. 8.4.1 show the online algorithm to be reasonably accurate for a wide range of δ in the binomial and Zipf/zeta case, *ii*) As observed in Sec. 8.4.3 for Fig. 8.3 and Fig. 8.4, there is a substantial difference in the sample size $\kappa(\delta)$ required between the binomial and Zipf/zeta dataset distributions for a given value of δ, e.g., 20% sample size for the binomial distribution vs. 70% sample size for the Zipf/zeta distribution to find an index with value within 90% of $\mathsf{X}_{n:n}$, *iii*) The right side of Fig. 8.5 shows $r(\tilde{k}(\delta))$ achieves close to the target ratio $1 - \delta$ for $\delta < 0.1$, in both the binomial and Zipf/zeta cases.

8.5 Related work

This section focuses on the problem of identifying near-maximum degree vertices in graphs using random sampling, ignoring random walk sampling, and only touching on EVT. *Random sampling:* Star sampling, a variation of snowball sampling Goodman [1961], Frank [1977], Kolaczyk [2009], can be effective in finding vertices of interest, as demonstrated in Chaps. 5 and 6. The key difference between this chapter and prior chapters is that formerly it has been assumed a target vertex may be immediately identified. In the context of finding maximum degree vertices, this assumes the maximum degree is known. The contribution of this chapter is to remove this assumption via

online estimation. *EVT for maximum degree search:* The "two-stage" graph sampling algorithm in Avrachenkov et al. [2014] is a nice application of extreme value theory to efficiently find the set of k vertices in a large directed graph with the largest in-degree. As the sampling methods presented in this chapter are not tied to graphs, while less efficient than the "two-stage" algorithm, they apply to a more general set of problems.

8.6 Conclusions

A method is proposed for circumventing the difficulty of finding a near-maximum value when the true maxima is unknown via online parameter estimation. The novelty is in using EVT to compute the expected required sample size for the sampled distribution. In identifying values within 90% of the maximum value, the proposed method is shown to yield reasonably accurate estimates in both binomial and Zipf/zeta dataset distributions.

Chapter 9: Conclusion

This thesis investigates the expected cost (e.g., number of queries) in using random walk and random sampling techniques to search large graphs to find a vertex with a particular property. Highlights of our findings include:

- Chap. 2 shows that, when searching for maximum-degree vertices, the cost of random walk techniques may be reduced, relative to the cost of star sampling, by selecting a suitable bias parameter given the assortativity.
- Chap. 3 shows that the asymptotic (in the graph order) expected number of maximum degree vertices in an Erdős Rényi (ER) graph is near one, i.e., there is typically a unique maximum degree vertex.
- Chap. 4 introduced the Self-Avoiding Walk Jump (SAWJ) algorithm for finding a maximum degree vertex, modeled its performance, and compared that performance to competing algorithms from the literature.
- Chap. 5 presented preliminary investigations into the expected number of samples required to find a target vertex when using star sampling with replacement.
- Chap. 6 gives analytical expressions for the probability and the approximate probability of the three variants of star sampling hitting the target set for the first time at sample t and the expected cost and approximate expected cost for three variants of star sampling under both the unit and linear cost models.
- Chap. 7 studies greedy graph wiring and greedy graph rewiring constructions intended to produce graphs with maximum assortativity over all graphs with a given degree sequence, and shows that these greedy approaches may fail.
- Chap. 8 analyzes the question of how many samples are needed in order to find a near maximum value in a long list of independent and identically distributed random variables if the distribution of values is know but the maximum value is unknown.

Appendix A: List of submitted and published papers

- J. Stokes and S. Weber. A Markov chain model for the search time for max degree nodes in a graph using a biased random walk. In Information Sciences and Systems (CISS), March 2016. (Published)
- J. Stokes and S. Weber. The self-avoiding walk-jump (SAWJ) algorithm for finding maximum degree nodes in large graphs. In 2016 IEEE International Conference on Big Data (Big Data), Dec 2016. (Published)
- J. Stokes and S. Weber. On random walks and random sampling to find max degree nodes in assortative Erds Rnyi graphs. In 2016 IEEE Global Communications Conference (GLOBECOM), Dec 2016. (Published)
- J. Stokes and S. Weber. On the number of star samples to find a vertex or edge with given degree in a graph. In Information Sciences and Systems (CISS), March 2017. (Published)
- J. Stokes and S. Weber. Star sampling with and without replacement. In Proceedings of the International Workshop on Mining and Learning with Graphs (MLG), August 2017. (Published)
- J. Stokes and S. Weber. Common greedy wiring and rewiring heuristics do not guarantee maximum assortative graphs of given degree. In Information Processing Letters (IPL). (To be published)
- J. Stokes and S. Weber. Online estimation for finding a near-maximum value in a large list of numerical data. In Information Sciences and Systems (CISS), March 2018 (Published)
- J. Stokes and S. Weber. Extension to SAWJ paper. (To be submitted to Journal of Internet Mathematics 2018)
- J. Stokes and S. Weber. Extension to KDD paper. (To be submitted to IEEE Transactions of Network Science 2018)

Bibliography

Jure Leskovec and Andrej Krevl. Stanford Large Network Dataset Collection (SNAP). `http://snap.stanford.edu/data`, June 2014.

M. E. J. Newman. Assortative mixing in networks. *Physics Review Letters*, 89(20):208701, October 2002.

Konstantin Avrachenkov, Nelly Litvak, Marina Sokol, and Don Towsley. *Quick Detection of Nodes with Large Degrees*, pages 54–65. Springer Berlin Heidelberg, 2012.

R. Taylor. Constrained switchings in graphs. In *Proceedings of the Eighth Australian Conference on Combinatorial Mathematics*, volume 884, pages 314–336. Springer Berlin Heidelberg, 1981.

P. Erdős and T. Gallai. Graphs with prescribed degrees of vertices (Hungarian). *Matematikai Lapok*, 11:264–274, 1960.

Rex K. Kincaid, Sarah J. Kunkler, Michael Drew Lamar, and David J. Phillips. Algorithms and complexity results for finding graphs with extremal randić index. *Wiley Networks*, 67(4):338–347, July 2016.

S. Ikeda and I. Kubo. Impact of local topological information on random walks on finite graphs. In *Proc. of the 30th Intl. Conf. on Automata, Languages and Programming*, pages 1054–1067, 2003.

Colin Cooper, Tomasz Radzik, and Yiannis Siantos. A fast algorithm to find all high-degree vertices in graphs with a power-law degree sequence. *Internet Mathematics*, 10(1-2):137–161, 2014.

A.S. Maiya and T.Y. Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proc. of the 17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 105–113, 2011.

J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Springer, 1st edition, 1983.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

S. Janson and T. Luczak. *Theory of random graphs*. John Wiley, 2000.

R. Xulvi-Brunet and I. Sokolov. Changing correlations in networks: Assortativity and dissortativity. *Acta Physica Polonica B*, 36(5):1431–1455, May 2005.

Bennett Eisenberg, Gilbert Stengle, and Gilbert Strang. The asymptotic probability of a tie for first place. *The Annals of Applied Probability*, 3(3):731–745, 1993.

J.J.A.M. Brands, F.W. Steutel, and R.J.G. Wilms. On the number of maxima in a discrete sample. *Elsevier Journal of Statistics and Probability Letters*, 20:209–217, 1994.

Y. Baryshnikov, B. Eisenberg, and G. Stengle. A necessary and sufficient condition for the existence of the limiting probability of a tie for first place. *Elsevier Journal of Statistics and Probability Letters*, 23:203–209, 1995.

Yongcheng Qi and R.J.G. Wilms. The limit behavior of maxima modulo one and the number of maxima. *Elsevier Statistics and Probability Letters*, 34:75–84, 1997.

Yongcheng Qi. A note on the number of maxima in a discrete sample. *Elsevier Statistics and Probability Letters*, 33:373–377, 1997.

Peter Olofsson. A Poisson approximation with applications to the number of maxima in a discrete sample. *Elsevier Statistics and Probability Letters*, 44:23–27, 1999.

F. Thomas Bruss and Rudolf Grübel. On the multiplicity of the maximum in a discrete sample. *The Annals of Applied Probability*, 13(4):1252–1263, 2003.

Bennett Eisenberg. On the expectation of the maximum of iid geometric random variables. *Elsevier Statistics and Probability Letters*, 78:135–143, 2008.

Bennett Eisenberg. The number of players tied for the record. *Elsevier Statistics and Probability Letters*, 79:283–288, 2009.

Colin Cooper, Tomasz Radzik, and Yiannis Siantos. A fast algorithm to find all high degree vertices in power law graphs. In *Proc. of the 21st International Conference on World Wide Web*, pages 1007–1016. ACM, 2012.

K. Avrachenkov, N. Litvak, L. O. Prokhorenkova, and E. Suyargulova. Quick detection of high-degree entities in large directed networks. In *2014 IEEE International Conference on Data Mining*, pages 20–29, 2014.

M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, Feb 2003.

Eric D. Kolaczyk. *Statistical analysis of network data : methods and models*. Springer, New York,, London, 2009.

Ian F. Blake and H. Darabian. Approximation for the probability in the tails of the binomial distribution. *IEEE Transactions on Information Theory*, IT-33(3):426–428, May 1987.

Zoltán Sasvári. Inequalities for binomial coefficients. *Acad. Press Journal of Mathematical Analysis and Applications*, 236:223–226, 1999.

C. Avin and B. Krishnamachari. The power of choice in random walks: an empirical study. *Computer Networks*, 52(1):44–60, 2008.

Chul-Ho Lee, Xin Xu, and Do Young Eun. Beyond random walk and Metropolis-Hastings samplers. *ACM Performance Evaluation Review*, 40(1):319–330, June 2012.

L. Jin, Y. Chen, P. Hui, C. Ding, T. Wang, A. Vasilakos, B. Deng, and X. Li. Albatross sampling: Robust and effective hybrid vertex sampling for social graphs. In *Proceedings of the 3rd ACM International Workshop on MobiArch*, HotPlanet '11, pages 11–16, 2011.

B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *ACM Internet Measurement Conference (IMC)*, pages 390–403, Melbourne, Australia, November 2010.

Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

Jerry Alan Veeh. The multivariate Laplace-De Moivre theorem. *Journal of Multivariate Analysis*, 18(1):46–51, 1986.

V. Bentkus. On the dependence of the berryesseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385 – 402, 2003.

K.W. Breitung. *Asymptotic Approximations for Probability Integrals*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2006.

Chris Chatfield and A.J. Collins. *Introduction to Multivariate Analysis*. CRC Texts in Statistical Science. Chapman & Hall, 1st edition, 1980.

J. Stokes and S. Weber. The self-avoiding walk-jump (SAWJ) algorithm for finding maximum degree nodes in large graphs. In *2016 IEEE International Conference on Big Data (Big Data)*, Dec 2016.

O. Frank. Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235 – 264, 1977.

L. Goodman. Snowball Sampling. *Ann. Math. Statist.*, 32(1):148–170, 1961.

GMP Van Kempen and LJ Van Vliet. Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry Part A*, 39(4):300–305, 2000.

Howard Seltman. Approximations for mean and variance of a ratio. `http://www.stat.cmu.edu/~hseltman/files/ratio.pdf`, May 2017.

Robert A. Lew. Bounds on negative moments. *SIAM Journal of Applied Mathematics*, 30(4): 728–731, June 1976.

Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *Nature*, 393 (6684):440–442, 1998.

Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. URL `http://networkrepository.com`.

Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73:016102, 2006.

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 835–844, 2007.

Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013.

H. Wang and J. Lu. Detect inflated follower numbers in OSN using star sampling. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 127–133, 2013.

J. Stokes and S. Weber. Star sampling with and without replacement. In *Proceedings of the International Workshop on Mining and Learning with Graphs (MLG)*, August 2017.

Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 631–636, 2006.

Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. Improving random walk estimation accuracy with uniform restarts. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 98–109. Springer, 2010.

E. Voudigari, N. Salamanos, T. Papageorgiou, and E. J. Yannakoudakis. Rank degree: An efficient algorithm for graph sampling. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 120–129, 2016.

Daniel Stutzbach, Reza Rejaie, Nick Duffield, Subhabrata Sen, and Walter Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Trans. Netw.*, 17(2):377–390, 2009.

R. H. Li, J. X. Yu, L. Qin, R. Mao, and T. Jin. On random walk based graph sampling. In *2015 IEEE 31st International Conf. on Data Engineering*, pages 927–938, 2015.

Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. Walking in Facebook: A case study of unbiased sampling of osns. In *Proceedings of the 29th Conference on Information Communications*, INFOCOM'10, pages 2498–2506, 2010.

Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: Stratified sampling via weighted random walks. In *Proc. of the ACM SIGMETRICS Joint Intl. Conf. on Measurement and Modeling of Computer Systems*, pages 281–292, 2011.

Flavio Chiericetti, Anirban Dasgupta, Ravi Kumar, Silvio Lattanzi, and Tamás Sarlós. On sampling nodes in a network. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 471–481. International World Wide Web Conferences, 2016.

Colin Cooper and Alan Frieze. The cover time of sparse random graphs. *Random Structures & Algorithms*, 30(1-2):1–16, 2007.

Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64:046135, 2001.

Qin Lv, Pei Cao, Edith Cohen, Kai Li, and Scott Shenker. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th International Conference on Supercomputing*, ICS '02, pages 84–95, 2002.

Christos Gkantsidis, Milena Mihail, and Amin Saberi. Random walks in peer-to-peer networks: Algorithms and evaluation. *Performance Evaluation*, 63(3):241 – 263, 2006.

Mickey Brautbar and Michael Kearns. Local algorithms for finding interesting individuals in large networks. In *ICS*, 2010.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 137–146, Washington, D.C., August 2003.

Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Physical Review Letters*, 86:3682–3685, 2001.

P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers. Influence of assortativity and degree-preserving rewiring on the spectra of networks. *The European Physical Journal B*, 76(4): 643–652, 2010.

Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simon, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltan Toroczkai, Marian Boguna, Guido Caldarelli, Santo Fortunato, and Dmitri Krioukov. Quantifying randomness in real networks. *Nature Communications*, 6:8627, 2015.

Isabelle Stanton and Ali Pinar. Constructing and sampling graphs with a prescribed joint degree distribution. *ACM Journal of Experimental Algorithmics*, 17:3.5:3.1–3.5:3.25, September 2012.

M. Gjoka, B. Tillman, and A. Markopoulou. Construction of simple graphs with a target joint degree matrix and beyond. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1553–1561, 2015.

H. J. Ryser. Combinatorial properties of matrices of zeros and ones. *Canadian Journal of Mathematics*, 9:371–377, 1957.

Ravi Kannan, Prasad Tetali, and Santosh Vempala. Simple markov-chain algorithms for generating bipartite graphs and tournaments. *Random Structures & Algorithms*, 14(4):293–308, 1999.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

W. Winterbach, D. de Ridder, H. J. Wang, M. Reinders, and P. Van Mieghem. Do greedy assortativity optimization algorithms produce good results? *The European Physical Journal B*, 85(5): 151, 2012.

Jin Zhou, Xiaoke Xu, Jie Zhang, Junfeng Sun, Micheal Small, and Jun-An Li. Generating an assortative network with a given degree distribution. *International Journal of Bifurcation and Chaos*, 18(11):3495–3502, 2008.

Natarajan Meghanathan. Maximal assortative matching for complex network graphs. *Journal of King Saud University – Computer and Information Sciences*, 28(2):230–246, 2016.

Brendan D. McKay and Adolfo Piperno. Practical graph isomorphism II. *Journal of Symbolic Computation*, 60:94–112, 2014.

B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2008.

Robert L. Wolpert. Duke University, Department of Statistical Science, Statistics 230 Lecture Notes: Extremes. `http://www2.stat.duke.edu/courses/Fall13/sta230/lec/extremes.pdf`, Fall 2013.

M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *Eur. Phys. J. B - Condensed Matter and Complex Systems*, 41(2):255–258, Sep 2004.