University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Mathematics and Statistics Department Faculty Publication Series

Mathematics and Statistics

2019

Alignment strength and correlation for graphs

Donniell E. Fishkind
Johns Hopkins University

Lingyao Meng
Johns Hopkins University

Ao Sun
Johns Hopkins University

Carey E. Priebe
Johns Hopkins University

Vince Lyzinski
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/math_faculty_pubs

Recommended Citation
Fishkind, Donniell E.; Meng, Lingyao; Sun, Ao; Priebe, Carey E.; and Lyzinski, Vince, "Alignment strength and correlation for graphs" (2019). *Pattern Recognition Letters*. 1295.
https://doi.org/10.1016/j.patrec.2019.05.008

This Article is brought to you for free and open access by the Mathematics and Statistics at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Mathematics and Statistics Department Faculty Publication Series by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Pattern Recognition Letters 125 (2019) 295–302

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Alignment strength and correlation for graphs☆

Donniell E. Fishkind[a,*], Lingyao Meng[a], Ao Sun[a], Carey E. Priebe[a], Vince Lyzinski[b]

[a] Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA
[b] Department of Mathematics and Statistics, University of Massachusetts Amherst, 710 N Pleasant Street Amherst, MA 01003, USA

ARTICLE INFO

Article history:
Received 13 November 2018
Available online 7 May 2019

MSC:
05C80
05C60
90C35

Keywords:
Correlated Bernoulli random graphs
Alignment strength
Graph correlation
Graph matchability
Complexity of graph matching

ABSTRACT

When two graphs have a correlated Bernoulli distribution, we prove that the alignment strength of their natural bijection strongly converges to a novel measure of graph correlation ϱ_T that neatly combines intergraph with intragraph distribution parameters. Within broad families of the random graph parameter settings, we illustrate that exact graph matching runtime and also matchability are both functions of ϱ_T, with thresholding behavior starkly illustrated in matchability.

© 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Overview

Suppose G and H are any two graphs with the same number of vertices. For any positive integer n, define $[n] := \{1, 2, 3, \ldots, n\}$, and let $\binom{[n]}{2}$ denote the set of all 2-element subsets of $[n]$. For simplicity, suppose that the vertex sets of G and H are both $[n]$. Let Π_n denote the set of bijections from $[n]$ to $[n]$. For each $\phi \in \Pi_n$, we define the *number of disagreements between G and H under* ϕ to be

$$d(G, H, \phi) := \sum_{\{i,j\} \in \binom{[n]}{2}} \mathbb{1}\Big(\mathbb{1}\big(i \sim_G j\big) \neq \mathbb{1}\big(\phi(i) \sim_H \phi(j)\big) \Big), \quad (1)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and $\sim_G$ denotes adjacency of vertices in G.

For each $\phi \in \Pi_n$, we define the *alignment strength of* ϕ as

$$\mathfrak{str}(G, H, \phi) := 1 - \frac{d(G, H, \phi)}{\frac{1}{n!} \sum_{\phi' \in \Pi_n} d(G, H, \phi')}. \quad (2)$$

The denominator in this definition of alignment strength serves as a normalizing factor; in particular, if ϕ is an isomorphism between G and H then the alignment strength of ϕ is 1, and if the number of adjacency disagreements for ϕ is merely average among the bijections in Π_n then the alignment strength of ϕ is 0. (Of course, if G and H are both edgeless or both complete graphs then $\mathfrak{str}(G, H, \phi)$ is not defined.)

If $\phi \in \Pi_n$ happens to be a known "natural alignment" between G and H (for example, if G and H are social networks with the same members, and ϕ maps each member to themselves; e.g. an email network and a Twitter network with the same users) then $\mathfrak{str}(G, H, \phi)$ can be viewed as a numerical measure of the structural similarity between G and H. However, if a natural alignment between G and H is not known, then we can use the *graph matching problem solution*, which is defined as $\phi_{GM} \in \arg\min_{\phi' \in \Pi_n} d(G, H, \phi')$; specifically, $\mathfrak{str}(G, H, \phi_{GM})$ can be viewed as a numerical measure of the structural similarity between G and H.

Two practical notes regarding computation: Although the denominator $\frac{1}{n!} \sum_{\phi' \in \Pi_n} d(G, H, \phi')$ in the definition of alignment strength (Eq. (2)) involves an exponentially sized summation, nonetheless it can be computed efficiently using Eq. (5) from Section 3. Also, although the computation of the graph matching problem solution ϕ_{GM} is intractable [4], nonetheless there are effective, efficient approximate graph matching algorithms that can be used [8,25], one of which we discuss and use later in this paper.

A brief outline of this paper is as follows.

In Section 2 we describe a very general random graph setting; G and H are random graphs with a correlated Bernoulli distribution. In particular, G and H share the same vertex set, and the identity bijection $\mathcal{I} \in \Pi_n$ is the natural alignment between G and H.

☆ Conflict of interest: None

* Corresponding author.
E-mail address: def@jhu.edu (D.E. Fishkind).

https://doi.org/10.1016/j.patrec.2019.05.008
0167-8655/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Each pair of vertices is assigned its own probability of adjacency ("Bernoulli parameter") in G and H, and the indicator Bernoulli random variable for adjacency of the pair in G and the indicator Bernoulli random variable for adjacency of the pair in H have Pearson correlation coefficient ϱ_e. Inherent to this model is the inter-graph (i.e. between G and H) statistical correlation ϱ_e and the intra-graph *heterogeneity correlation* parameter ϱ_h, which is a function of the Bernoulli coefficients that measures their variation. Then we define the key parameter ϱ_T as $1-\varrho_T := (1-\varrho_e)(1-\varrho_h)$; we call ϱ_T the *total correlation*.

In Section 3 we state and prove our main theoretical result, Theorem 4, which asserts that for G and H with a correlated Bernoulli distribution we have that the alignment strength of the identity bijection $\mathfrak{str}(G,H,\mathcal{I})$ is asymptotically equal to the total correlation parameter ϱ_T. This suggests that the total correlation ϱ_T is a meaningful measure of the structural similarity between the graphs G and H realized from the correlated Bernoulli distribution. Of note is that the total correlation is nicely and cleanly partitioned by the defining formula $1-\varrho_T = (1-\varrho_e)(1-\varrho_h)$; this illustrates a symmetry in the effect of (inter-graph parameter) edge correlation ϱ_e and the affect of (intra-graph parameter) heterogeneity correlation ϱ_h.

The subsequent sections, Sections 4 and 5, follow up with empirical illustrations that total correlation ϱ_T is a meaningful measure. As we vary the edge correlation ϱ_e together with the heterogeneity correlation ϱ_h for correlated Bernoulli graphs G and H in broad families of parameter settings, it turns out that the value of ϱ_T dictates (in Section 4) how successful the approximate seeded graph matching algorithm called SGM [8,16] is in recovering the identity bijection (which is the natural alignment here) and (in Section 5) ϱ_T dictates how much time it takes to perform seeded graph matching exactly via binary integer linear programming. The *seeded graph matching problem* is the graph matching problem wherein we seek to compute $\phi_{GM} \in \arg\min_{\phi' \in \Pi_n} d(G,H,\phi')$, except that part of the natural alignment is known; having these "seeds" can substantially help recover the rest of the natural alignment correctly. In Section 4, we utilize the SGM Algorithm [8,16] for approximate seeded graph matching on moderately sized graphs, on the order of 1000 vertices, since, unfortunately, exact seeded graph matching can only be done on very small, toy-size graphs (a few tens of non-seed vertices). In Section 5, where we are interested in comparing runtime, the approximate seeded graph matching algorithms are not appropriate to use, since their run times tend to be monolithic (given the number of vertices) and less sensitive to the parameters of the random graph distribution. So we do exact seeded graph matching, but only on small enough examples.

2. Random graph setting: correlated Bernoulli graphs

In this section we describe the correlated Bernoulli random graph distribution, and three important associated parameters/functions of parameters; namely ϱ_e, ϱ_h, and ϱ_T.

For any positive integer n, the distribution parameters are any given real number ϱ_e (called the *edge correlation*) from the interval [0,1], and any given set of real numbers $\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}}$ (called the *Bernoulli parameters*) from the interval [0,1] such that the Bernoulli parameters are not all equal to 0 and not all equal to 1. Random graphs G and H, each on vertex set $[n]$, will be called *ϱ_e-correlated random Bernoulli*$(\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}})$ *graphs* if, for each $\{i,j\} \in \binom{[n]}{2}$, we have that $\mathbb{1}(i \sim_G j)$ is a Bernoulli$(p_{i,j})$ random variable, and $\mathbb{1}(i \sim_H j)$ is a Bernoulli$(p_{i,j})$ random variable, and, if $0 < p_{i,j} < 1$, then the two random variables $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ have Pearson correlation coefficient ϱ_e; other than these specified dependencies, the random variables $\{\mathbb{1}(i \sim_G j)\}_{\{i,j\}\in\binom{[n]}{2}} \bigcup \{\mathbb{1}(i \sim_H j)\}_{\{i,j\}\in\binom{[n]}{2}}$ are collectively independent.

Such G, H can be realized from this distribution as follows. For all $\{i,j\} \in \binom{[n]}{2}$ independently, first realize $\mathbb{1}(i \sim_G j)$ from the Bernoulli$(p_{i,j})$ distribution. Then, conditioned on $\mathbb{1}(i \sim_G j)$, realize $\mathbb{1}(i \sim_H j)$ from distribution Bernoulli$(\varrho_e \cdot \mathbb{1}(i \sim_G j) + (1-\varrho_e)\cdot p_{i,j})$. It is easy to verify that $\mathbb{1}(i \sim_H j)$ has a marginal distribution Bernoulli$(p_{i,j})$ and, indeed, the random variables $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ have Pearson correlation ϱ_e if $0 < p_{i,j} < 1$. Moreover, it easy to verify that, for any two Bernoulli$(p_{i,j})$ random variables such that $0 < p_{i,j} < 1$, the Pearson correlation coefficient uniquely determines their joint distribution. Also, it is easy to verify that $\mathbb{P}[i \sim_G j \;\&\; i \not\sim_H j] = (1-\varrho_e)p_{i,j}(1-p_{i,j})$. See Appendix A for more of all these details.

The identity bijection $\mathcal{I} \in \Pi_n$ is the natural alignment between G and H. When $\varrho_e = 1$ we have that G, H are almost surely isomorphic (via isomorphism $\mathcal{I}$), and when $\varrho_e = 0$ we have that G and H are independent (i.e. the indicators for all edges of both graphs are collectively independent). If all Bernoulli parameters $p_{i,j}$ are equal to each other then G and H are Erdos–Renyi random graphs.

Associated with the Bernoulli parameters $\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}}$, denote their mean

$$\mu := \frac{1}{\binom{n}{2}} \sum_{\{i,j\}\in\binom{[n]}{2}} p_{i,j}$$

and denote their variance

$$\sigma^2 := \frac{1}{\binom{n}{2}} \sum_{\{i,j\}\in\binom{[n]}{2}} (p_{i,j}-\mu)^2.$$

We define the *heterogeneity correlation* ϱ_h

$$\varrho_h := \frac{\sigma^2}{\mu(1-\mu)}. \tag{3}$$

It is simple to show that $0 \le \varrho_h \le 1$. Furthermore, $\varrho_h = 0$ if and only if all Bernoulli parameters $p_{i,j}$ are equal to each other (i.e. G and H are Erdos–Renyi random graphs), and $\varrho_h = 1$ if and only if all Bernoulli parameters are 0 or 1 (but, recall, the Bernoulli parameters are not all 0 and are not all 1). See Appendix B for more details. Note that ϱ_h is a measure of heterogeneity within G (and within H) by virtue of its numerator being the variance (a measure of spread) of the Bernoulli coefficients, although this variance is normalized through division by the denominator of ϱ_h, where this denominator is a function of the global graph density. (So, among distributions with a common global density, ϱ_h is just a multiple of the variance σ^2.)

Note that edge correlation ϱ_e is an inter-graph affect (between G and H), whereas heterogeneity correlation ϱ_h is an intra-graph affect. Unlike edge correlation ϱ_e, heterogeneity correlation ϱ_h is not a statistical correlation. However, our results will demonstrate that ϱ_h is interchangeable with edge correlation ϱ_e with regard to creating alignment strength. We thus take the liberty of calling ϱ_h "correlation," but we do so in a looser, nonstatistical sense, with the meaning that it generates similarity between G and H just like edge correlation does.

Finally, define the *total correlation* ϱ_T such that ϱ_T satisfies

$$1-\varrho_T := (1-\varrho_e)(1-\varrho_h). \tag{4}$$

3. Alignment strength is total correlation, asymptotically

In this section we state and prove our main theoretical result, Theorem 4, that when G, H have a correlated Bernoulli distribution then the identity bijection $\mathcal{I} \in \Pi_n$ (the natural alignment here) has alignment strength asymptotically equal to the distribution's total correlation ϱ_T.

Let e_G and e_H denote the number of edges in G and H, respectively, and let $\mathfrak{d}_G := \frac{e_G}{\binom{n}{2}}$ and $\mathfrak{d}_H := \frac{e_H}{\binom{n}{2}}$ respectively denote the densities of G and H.

Lemma 1. *For any graphs G, H on common vertex set $[n]$, and any $\phi \in \Pi_n$, it holds that*

$$\mathfrak{stt}(G,H,\phi) = 1 - \frac{\frac{d(G,H,\phi)}{\binom{n}{2}}}{\mathfrak{d}_G(1-\mathfrak{d}_H) + (1-\mathfrak{d}_G)\mathfrak{d}_H}.$$

Proof. With G and H fixed, consider random $\varphi \in \Pi_n$ with a discrete-uniform distribution; the expected value of $d(G, H, \varphi)$ is $\frac{1}{n!}\sum_{\phi' \in \Pi_n} d(G,H,\phi')$. We next equivalently compute the expected value of $d(G, H, \varphi)$ using linearity of expectation over the sum of its indicators in Eq. (1). Observe that, for any two vertices that form an edge in G, the probability that φ maps them to a nonedge of H is $\frac{\binom{n}{2}-e_H}{\binom{n}{2}}$, and, for any two nonadjacent vertices of G, the probability that φ maps them to an edge of H is $\frac{e_H}{\binom{n}{2}}$; the expected value of $d(G, H, \varphi)$ is thus

$$\begin{aligned}
&\frac{1}{n!}\sum_{\phi' \in \Pi_n} d(G,H,\phi') \\
&= e_G \cdot \frac{\binom{n}{2} - e_H}{\binom{n}{2}} + \left(\binom{n}{2} - e_G\right) \cdot \frac{e_H}{\binom{n}{2}} \\
&= \binom{n}{2} \cdot \left[\mathfrak{d}_G(1-\mathfrak{d}_H) + (1-\mathfrak{d}_G)\mathfrak{d}_H\right]. \qquad (5)
\end{aligned}$$

The desired result then follows from substituting Eq. (5) into the definition of $\mathfrak{stt}(G, H, \phi)$ in Eq. (2). □

In the rest of this section we will state and prove limit results for random correlated Bernoulli graphs G, H. This context requires us to consider a sequence of experiments —for each value of $n = 1, 2, 3, \ldots$ —wherein the chosen edge correlation ϱ_e is a function of n, and the chosen Bernoulli parameters $\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}}$ are also functions of n, and thus ϱ_h and ϱ_T are also functions of n. For ease of notation, we do not explicitly write argument n in these functions. However, we will require that there exists a positive lower bound for μ over all n, and as well that there exists an upper bound less than 1 for μ over all n. (Note that since μ is a function of n, we have that the μ are a sequence, so the following limit result is expressed as a difference that converges as stated, rather than convergence to μ, which would not make technical sense. Similarly for the other results here.)

Lemma 2. *We have $\mathfrak{d}_G - \mu \xrightarrow{a.s.} 0$ and $\mathfrak{d}_H - \mu \xrightarrow{a.s.} 0$.*

Proof. Clearly $\mathbb{E}(\mathfrak{d}_G) = \mu$. Also, e_G is the sum of $\binom{n}{2}$ independent Bernoulli random variables, and thus its variance is bounded by $\binom{n}{2}$, thus the variance of $\mathfrak{d}_G := \frac{e_G}{\binom{n}{2}}$ is of order $O(n^{-2})$. Next, by Chebyshev's Inequality, for any $\epsilon > 0$, $\mathbb{P}[|\mathfrak{d}_G - \mu| \geq \epsilon] \leq \frac{1}{\epsilon^2}\mathrm{Var}(\mathfrak{d}_G)$; since this probability is $O(n^{-2})$ when ϵ is fixed, it has finite sum over $n = 1, 2, 3, \ldots$. Thus, since ϵ is arbitrary, by the Borel–Cantelli Theorem $\mathfrak{d}_G - \mu \xrightarrow{a.s.} 0$, as desired. □

Theorem 3. *We have $\frac{d(G,H,\mathcal{I})}{\binom{n}{2}} - 2(1-\varrho_e)\left(\mu(1-\mu) - \sigma^2\right) \xrightarrow{a.s.} 0$*

Proof. We begin by taking the expected value of $d(G, H, \mathcal{I})$;

$$\begin{aligned}
&\mathbb{E}\left[d(G,H,\mathcal{I})\right] \\
&= \mathbb{E}\left[\sum_{\{i,j\}\in\binom{[n]}{2}} \mathbb{1}\Big(\ \mathbb{1}\big(i \sim_G j\big) \ \neq\ \mathbb{1}\big(i \sim_H j\big)\ \Big)\right] \\
&= \sum_{\{i,j\}\in\binom{[n]}{2}} 2(1-\varrho_e)p_{i,j}(1-p_{i,j}) \\
&= 2(1-\varrho_e)\binom{n}{2}\Big(\mu(1-\mu) - \sigma^2\Big), \qquad (6)
\end{aligned}$$

thus $\mathbb{E}[\frac{d(G,H,\mathcal{I})}{\binom{n}{2}}] = 2(1-\varrho_e)(\mu(1-\mu) - \sigma^2)$.

Next, $d(G, H, \mathcal{I})$ is the sum of $\binom{n}{2}$ independent Bernoulli random variables, and thus its variance is bounded by $\binom{n}{2}$, thus the variance of $\frac{d(G,H,\mathcal{I})}{\binom{n}{2}}$ is of order $O(n^{-2})$. Next, by Chebyshev's Inequality, for any $\epsilon > 0$, $\mathbb{P}[|\frac{d(G,H,\mathcal{I})}{\binom{n}{2}} - 2(1-\varrho_e)(\mu(1-\mu) - \sigma^2)| \geq \epsilon] \leq \frac{1}{\epsilon^2}\mathrm{Var}(\frac{d(G,H,\mathcal{I})}{\binom{n}{2}})$; since this probability is $O(n^{-2})$ when ϵ is fixed, it has finite sum over $n = 1, 2, 3, \ldots$. Thus, since ϵ is arbitrary, by the Borel-Cantelli Theorem $\frac{d(G,H,\mathcal{I})}{\binom{n}{2}} - 2(1-\varrho_e)(\mu(1-\mu) - \sigma^2) \xrightarrow{a.s.} 0$, as desired. □

The following is the main result of this section, and is our main theoretical result.

Theorem 4. *It holds that $\mathfrak{stt}(G, H, \mathcal{I}) - \varrho_T \xrightarrow{a.s.} 0$*

Proof. By Lemma 2, $\mathfrak{d}_G - \mu \xrightarrow{a.s.} 0$ and $\mathfrak{d}_H - \mu \xrightarrow{a.s.} 0$. Because $\mathfrak{d}_G$, $\mathfrak{d}_H$ and μ are bounded, we thus have that $\mathfrak{d}_G(1-\mathfrak{d}_H) + (1-\mathfrak{d}_G)\mathfrak{d}_H - 2\mu(1-\mu) \xrightarrow{a.s.} 0$. Now, by Theorem 3, we have that $\frac{d(G,H,\mathcal{I})}{\binom{n}{2}} - 2(1-\varrho_e)\Big(\mu(1-\mu) - \sigma^2\Big) \xrightarrow{a.s.} 0$; since the relevant sequences are bounded, and μ is bounded away from 0 and 1, we have that

$$\frac{\frac{d(G,H,\mathcal{I})}{\binom{n}{2}}}{\mathfrak{d}_G(1-\mathfrak{d}_H) + (1-\mathfrak{d}_G)\mathfrak{d}_H} - \frac{2(1-\varrho_e)\Big(\mu(1-\mu) - \sigma^2\Big)}{2\mu(1-\mu)} \xrightarrow{a.s.} 0.$$

Applying Lemma 1 and the definitions of ϱ_h and ϱ_T we thus have from above that $(1 - \mathfrak{stt}(G, H, \mathcal{I})) - (1 - \varrho_T) \xrightarrow{a.s.} 0$, which proves Theorem 4. □

4. Graph matchability and total correlation ϱ_T

In this section we empirically demonstrate in broad families of parameter settings where ϱ_e and ϱ_h vary, that success of an approximate seeded graph matching algorithm is a function of ϱ_T.

Our setting is where G, H are correlated Bernoulli graphs on vertex set $[n]$. The graph matching problem is to compute $\phi_{GM} \in \arg\min_{\phi\in\Pi_n} d(G, H, \phi)$. In the *seeded graph matching problem*, there are s *seeds*, without loss of generality they are the vertices $1, 2, \ldots, s$, and there are $m := n - s$ ambiguous vertices, which are the other vertices $s+1, s+2, \ldots, n$. The meaning of *seeded graph matching* is that the feasible region $\phi \in \Pi_n$ of the graph matching problem is restricted to $\phi \in \Pi_n$ that satisfy $\phi(i) = i$ for all seeds $i = 1, 2, \ldots, s$. The graphs G and H are separately observed and the identities of the ambiguous vertices are unobserved for the optimization, so that the natural alignment, which is the identity bijection $\mathcal{I}$, is only seen for the seeds. If the seeded graph matching solution is $\mathcal{I}$ then we say that G and H are *matchable*.

Even a modest number of seeds can make a very significant increase in the likelihood that G and H are matchable [16]. Our illustration in this section will be for realistically sized graphs, on the order of a thousand vertices, and we utilize seeds because they will be quite helpful in obtaining reasonable probability of matchability. Unfortunately, exact graph matching –even seeded graph matching– is intractable, only solvable on the smallest, toy examples. So we utilize an approximate seeded graph matching algorithm; the specific one we use is the SGM Algorithm [8,16], which has been demonstrated to have many nice theoretical properties,

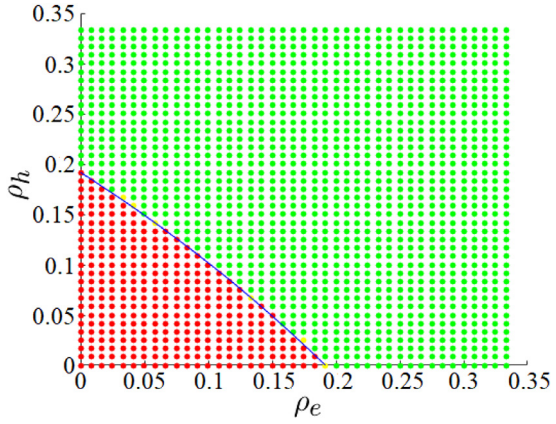


Fig. 1. Matchability experiment for $m = 850$, $s = 150$, $p = \frac{1}{2}$.

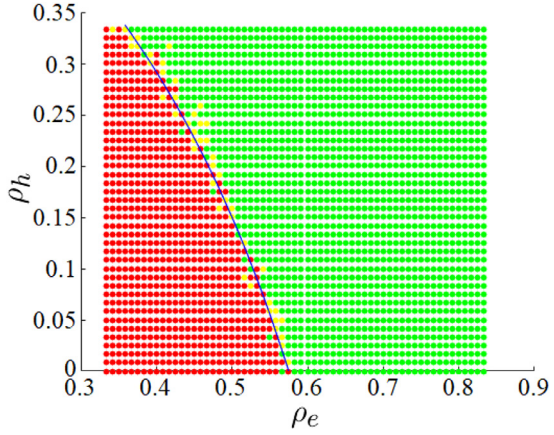


Fig. 2. Matchability experiment for $m = 850$, $s = 9$, $p = \frac{1}{2}$.

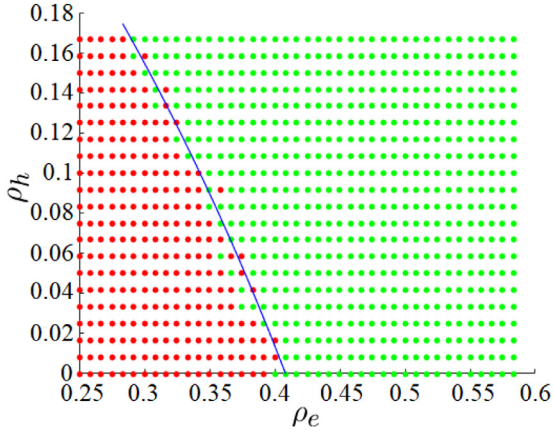


Fig. 3. Matchability experiment for $m = 850$, $s = 22$, $p = \frac{1}{3}$.

and it is efficient and quite effective (see [8,15,16]). In this section, we will say that G and H are matchable if the SGM-generated approximate seeded graph matching solution is the identity bijection $\mathcal{I}$.

In the experiments that we will perform, we will sample G, H from a correlated Bernoulli distribution for different values of ϱ_e and ϱ_h; the values of the Bernoulli coefficients $\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}}$ are selected as follows, in order to obtain specified values of ϱ_h. Given any real number $p \in (0, 1)$ and real number $\delta \in [0, \min\{p, 1-p\}]$, we independently randomly sample $\{p_{i,j}\}_{\{i,j\}\in\binom{[n]}{2}}$ from the uniform distribution on the interval $(p-\delta, p+\delta)$. Note that the aforedefined Bernoulli parameter variance σ^2 has expected value $\frac{\delta^2}{3}$, and σ^2 will be approximately $\frac{\delta^2}{3}$ for large values of n. For a fixed p, as δ goes from 0 to $\min\{p, 1-p\}$, the value of $\varrho_h = \frac{\sigma^2}{\mu(1-\mu)} \approx \frac{\delta^2}{3p(1-p)}$ monotonically increases from 0 to $\frac{1}{3}\cdot\frac{1-p}{p}$ if $p \geq \frac{1}{2}$ and $\frac{1}{3}\cdot\frac{p}{1-p}$ if $p \leq \frac{1}{2}$. In this section and in the next section, when we report values of ϱ_e and ϱ_h, we mean that we selected δ so that the approximate value of ϱ_h is as reported.

We did three batches of experiments. In the first batch of experiments, for each value of $\varrho_e = 0, \frac{1}{120}, \frac{2}{120}, \frac{3}{120}, \ldots, \frac{1}{3}$ and $\varrho_h = 0, \frac{1}{120}, \frac{2}{120}, \frac{3}{120}, \ldots, \frac{1}{3}$, we did 60 replicates of obtaining random graphs G, H with $m = 850$ ambiguous vertices and $s = 150$ seeds from a correlated Bernoulli distribution with edge correlation ϱ_e and heterogeneity correlation ϱ_h based on $p = \frac{1}{2}$, and we performed seeded graph matching with the SGM algorithm. If all 60 replicates were matchable then we plotted a green dot in Fig. 1 at the appropriate coordinates, if between 1 and 5 of the 60 replicates were not matchable then we plotted a yellow dot in the figure, and if more than 5 of the 60 replicates were not matchable then we plotted a red dot. The blue curve in the figure is the set of all pairs of ϱ_e, ϱ_h such that $\varrho_T = \frac{23}{120}$.

In these experiments and those below, the transition from matchable to anonymized (i.e., not matchable) occurs at a level set of ϱ_T. We note here that numerous results in the literature have studied this matchability phase transition as a function of edge correlation ϱ_e (see, for example, [5,6,16]) and a few papers have considered the impact of network heterogeneity on matchability (see, for example, [14,18]). In the parameterized correlated Bernoulli distribution considered above, these empirical results novelly suggest the form by which matchability is impacted by within and across graph correlation structure. Further understanding this phase transition as a function of ϱ_T is a necessary next step to understand the dual roles that graph structure (ϱ_h) and graph pairedness (ϱ_e) play in network alignment problems both theoretical and practical.

The second batch of experiments differed just in that there were only $s = 9$ seeds (with $m = 850$ as before), and the range of values of ϱ_e was $\frac{1}{3}$ to $\frac{5}{6}$ in increments of $\frac{1}{120}$; the results are similarly displayed in Fig. 2, and the blue curve in the figure is the set of all pairs of ϱ_e, ϱ_h such that $\varrho_T = \frac{69}{120}$. In these experiments, we again see the transition in matchability at a level set of ϱ_T, although the transition is looser due to fewer seeds being considered in this problem setup.

The third batch of experiments differed just in that there were $s = 22$ seeds, and now $p = \frac{1}{3}$, the range of values of ϱ_e was $\frac{1}{4}$ to $\frac{7}{12}$ in increments of $\frac{1}{120}$, and the range of values of ϱ_h was 0 to $\frac{1}{6}$ in increments of $\frac{1}{120}$; the results are similarly displayed in Fig. 3, and the blue curve in the figure is the set of all pairs of ϱ_e, ϱ_h such that $\varrho_T = \frac{49}{120}$. In these experiments, we again see the transition in matchability at a level set of ϱ_T.

We then repeated the above experiments for each combination of: total number of vertices 300 or 600, number of seeds 5% or 10% of the vertices, and values of p being $\frac{1}{2}$ or $\frac{1}{3}$. In all eight such combinations the result of the experiments were like the above; namely, matchability was a function of ϱ_T.

Note that matchability is not universally a function of just ϱ_T. For example, the number of vertices and the number of seeds have a dramatic effect on matchability. The empirical demonstrations in this section of matchability as a function of ϱ_T are limited to families of correlated Bernoulli distribution parameterizations of the type that we have used here. New work will be needed to obtain theorems that universally and fully account for matchability. But, nonetheless, we have empirically demonstrated in broad families of parameter settings that the phase transition in matchability

occurs at a level set of ϱ_T, which supports the importance and utility of ϱ_T as a meaningful measure of graph correlation.

5. Graph matching runtime and total correlation ϱ_T

Similar to the previous section, in this section we empirically demonstrate, in broad families of parameter settings where ϱ_e and ϱ_h vary, that the running time of exact seeded graph matching via binary integer linear programming is a function of ϱ_T.

We consider exact seeded graph matching here because the approximate seeded graph matching algorithms have running times that are relatively monolithic (when the number of vertices are fixed) and not sensitive enough to the parameters in the random graph distribution. Unfortunately, exact graph matching is intractable [4], and can only be done for small examples; we will work with graphs that have 20 ambiguous vertices.

For this section, the random graphs G, H have correlated Bernoulli distributions, for various values of ϱ_e and ϱ_h. The Bernoulli parameters are chosen in exactly the manner of the previous section, Section 4; there is a fixed value p, and then δ are selected to attain desired values of ϱ_h in the manner described in the previous section.

We next formulate the binary integer linear program for seeded graph matching. For graphs G and H, say their adjacency matrices are A and B, respectively, and say that there are s seeds and m ambiguous vertices. We partition $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ and $B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$, where A_{11}, $B_{11} \in \{0,1\}^{s\times s}$, A_{12}, $B_{12} \in \{0,1\}^{s\times m}$, A_{21}, $B_{21} \in \{0,1\}^{m\times s}$, and A_{22}, $B_{22} \in \{0,1\}^{m\times m}$. (Note that $A_{12} = A_{21}^T$ and $B_{12} = B_{21}^T$ here, since A and B are symmetric, but we do not use this fact in the formulation below so that the formulation is expressed even more generally.) Let I denote the identity matrix (subscripted with its number of rows and columns), let 0 subscripted denote the matrix of zeros of subscripted size, let $\vec{1}$ denote the column vector of ones with subscripted number of entries, let $\vec{0}$ denote the column vector of zeros with subscripted number of entries, let $\otimes$ denote the Kronecker product of matrices, let $\|\cdot\|_1$ denote the ℓ_1 vector norm for matrices (this norm is evaluated by taking the sum of absolute values of the matrix entries), for any matrix N let vec$N$ denote the column vector which is the concatenation of the columns of N (first column of N, then second column of N, etc., then last column of N), and let $\mathcal{P}_m$ denote the set of $m \times m$ permutation matrices. Clearly, the seeded graph matching problem is $\min_{P\in\mathcal{P}_m} \|A - \begin{bmatrix} I_{s\times s} & 0_{s\times m} \\ 0_{m\times s} & P \end{bmatrix} B \begin{bmatrix} I_{s\times s} & 0_{s\times m} \\ 0_{m\times s} & P \end{bmatrix}^T \|_1$. By permuting columns of the matrix in the norm, we get an equivalent formulation of the seeded graph matching problem as:

$$\min_{P\in\mathcal{P}_m} \|A \begin{bmatrix} I_{s\times s} & 0_{s\times m} \\ 0_{m\times s} & P \end{bmatrix} - \begin{bmatrix} I_{s\times s} & 0_{s\times m} \\ 0_{m\times s} & P \end{bmatrix} B \|_1.$$

Expanding this, we get an equivalent formulation of the seeded graph matching problem as:

$$\min_{P\in\mathcal{P}_m} \left(\|A_{12}P - B_{12}\|_1 + \|A_{21} - PB_{21}\|_1 + \|A_{22}P - PB_{22}\|_1 \right). \tag{7}$$

Now, because of the absolute values in $\|\cdot\|_1$, we add artificial variables to obtain simple linearity. For example, (just) minimizing $\|A_{22}P - PB_{22}\|_1$ subject to $P \in \mathcal{P}_m$ is equivalent to minimizing the sum of the entries of $E + E'$ subject to $A_{22}P - PB_{22} + E - E' = 0_{m\times m}$, $P \in \mathcal{P}_m$, E, $E' \in \{0,1\}^{m\times m}$. Of course, there are additional $\|\cdot\|_1$ terms in the objective function in Eq. (7), but the same approach can be used, so that seeded graph matching is equivalent to

$$\min \begin{bmatrix} \vec{0}_{m^2} \\ \vec{1}_{2m^2+4ms} \end{bmatrix}^T x$$

$$\text{s.t. } [M|E]x = b$$

$$x \in \{0,1\}^{3m^2+4ms}$$

where the first m^2 entries of x are vec$P$, and M and E and b are given by:

$$M = \begin{bmatrix} I_{m\times m} \otimes A_{22} - B_{22}^T \otimes I_{m\times m} \\ I_{m\times m} \otimes A_{12} \\ B_{21}^T \otimes I_{m\times m} \\ I_{m\times m} \otimes \vec{1}_m^T \\ \vec{1}_m^T \otimes I_{m\times m} \end{bmatrix}$$

$$E = \begin{bmatrix} I_{(m^2+2ms)\times(m^2+2ms)} & -I_{(m^2+2ms)\times(m^2+2ms)} \\ 0_{2m\times(m^2+2ms)} & 0_{2m\times(m^2+2ms)} \end{bmatrix}$$

$$b = \begin{bmatrix} \vec{0}_{m^2} \\ \text{vec}B_{12} \\ \text{vec}A_{21} \\ \vec{1}_m \\ \vec{1}_m \end{bmatrix}$$

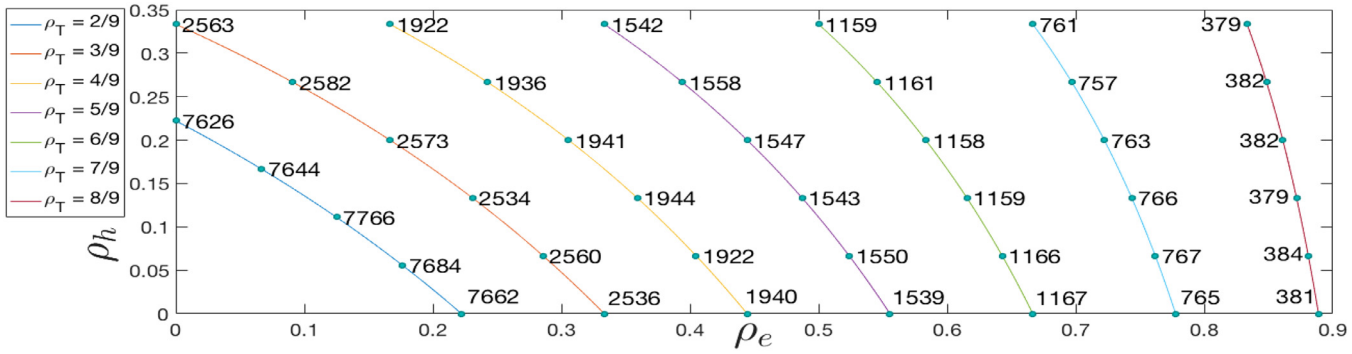
We solve the above binary integer linear program exactly using the optimization package GUROBI. The yardstick for runtime that we have chosen to adopt is the number of simplex iterations performed by GUROBI; this measure has the advantage of reducing many sources of platform variability.

We performed three batches of experiments. In the first batch of experiments, for each value of $\varrho_T = \frac{2}{9}, \frac{3}{9}, \frac{4}{9}, \ldots, \frac{8}{9}$, we selected various pairs of ϱ_e, ϱ_h which have $1 - \varrho_T = (1-\varrho_e)(1-\varrho_h)$ for the given value of ϱ_T; the values of ϱ_h are achieved based on $p = \frac{1}{2}$, and the chosen pairs ϱ_e, ϱ_h are the points plotted with a dot in Fig. 4a. For each such pair ϱ_e, ϱ_h we did 60 replicates of obtaining random graphs G, H with $m = 20$ ambiguous vertices and $s = 480$ seeds from a correlated Bernoulli distribution with edge correlation ϱ_e and heterogeneity correlation ϱ_h, and we solved the seeded graph matching problem for G, H exactly using GUROBI. The average runtimes (measured by the number of simplex iterations performed by GUROBI) are printed above each pair ϱ_e, ϱ_h at the appropriate coordinates in Fig. 4a. The smooth curves on the plot are the level sets of ϱ_T.

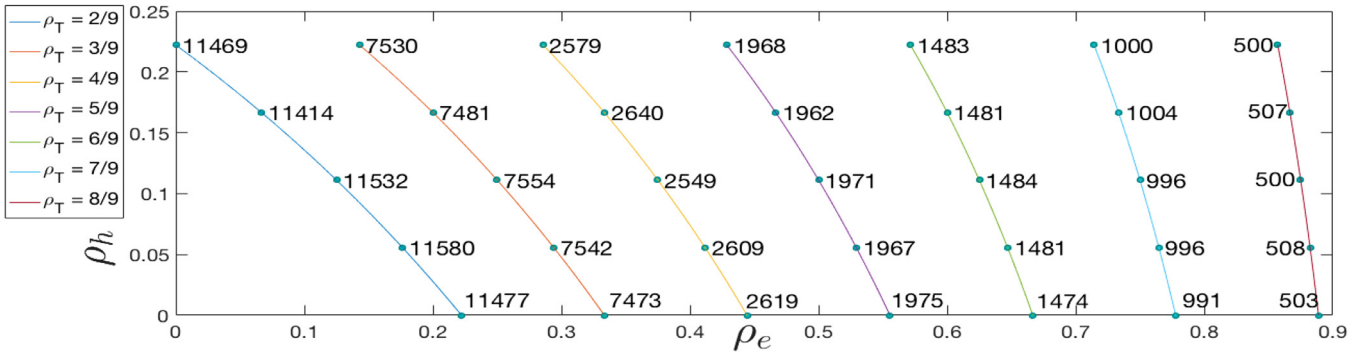
These experiments, and those below, suggest that in this parametrized Bernoulli graph model the algorithmic runtimes are approximately constant on the level sets of ϱ_T. The results in Section 4 suggest that the phase transition of matchability occurs at a level set of ϱ_T, and these results further reinforce the novel overarching notion: that the theoretic and algorithmic difficulty of matching is a function of ϱ_e and ϱ_h only through ϱ_T. Alone, ϱ_e and ϱ_h are insufficient to capture this theoretic and algorithmic difficulty.

The second and third batch of experiments are exactly like the first batch, except that for the second batch of experiments the values of ϱ_h are based on $p = \frac{3}{5}$ and the results are displayed in Fig. 4b, and for the third batch of experiments the values of ϱ_h are based on $p = \frac{1}{3}$ and the results are displayed in Fig. 4c. Note that the ranges of ϱ_h are different in Fig. 4a–c because different values of p put different limitations on δ.

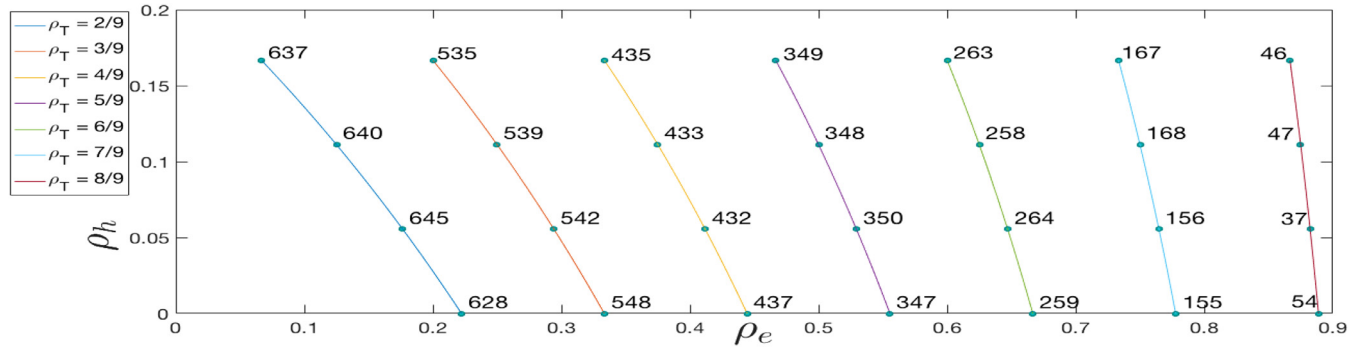
Just like for matchability in the previous section, it must be pointed out that the runtime of exact seeded graph matching via binary integer linear programming is not universally a function of ϱ_T. Of course, the number of vertices —particularly the number of ambiguous vertices— has a dominant role in the runtime, and the above experiments show that the graph density likewise plays a very large role. Nonetheless, for families of correlated Bernoulli distributed graphs similar to the ones in the experiments above, we see within a family that the runtime is a function of ϱ_T.



(a) Runtime experiment for $m = 20$, $s = 480$, $p = \frac{1}{2}$.



(b) Runtime experiment for $m = 20$, $s = 480$, $p = \frac{3}{5}$.



(c) Runtime experiment for $m = 20$, $s = 480$, $p = \frac{1}{3}$.

Fig. 4. Runtime experiments for different settings.

6. Discussion and future work

The correlated Bernoulli random graph model considered herein contains many important families of random graph models as subfamilies including stochastic blockmodels [1,11], random dot product graphs [2,27], and more general latent position random graph [10]. While the edge independent assumption inherent to these models is often not satisfied in real data applications, nonetheless (conditionally) edge-independent random models have shown great utility in capturing statistically relevant structure in a host of real data applications from modeling connectomic structure [13,17,20], to capturing community and user-level behavior in social networks [19,26]. Moreover, these models provide a theoretically tractable environment in which to explore important statistical concepts such as estimation consistency [3,21,22], consistent hypothesis testing [12,23,24], and network de-anonymization [6,7], among others. Indeed, it is this appealing mix of theoretical tractability and practical utility that have made these graph models an increasingly popular option in the statistical network inference community.

In this paper we prove in a very broad random graph setting—specifically, when G and H have a correlated Bernoulli distribution—that the alignment strength of the natural G, H alignment is asymptotically equal to the total correlation ϱ_T in the distribution. After this, we empirically demonstrate, for types of families within the distribution, that both matchability and exact-solution-runtime for seeded graph matching of G, H are functions of the total correlation ϱ_T.

Graph matching and seeded graph matching are extremely important in many disciplines; see the surveys [4] and [9]. Unfortunately, these problems are intractable; in their full generality they are NP-hard. Obtaining a function of the distribution parameters that universally predicts matchability via approximate algorithms would be a huge advance in theoretical understanding and

in practice. Likewise, it would be a huge advance to predict exact-solution-runtime from a function of the distribution parameters, and it would not just be the number of vertices—the other parameters play a large role. The goals of obtaining these universal functions has not been achieved here; the families we use here are general but not universal. But a universal result will include our families as special cases, thus ϱ_T will play an important role.

There are a number of matchability results already known, see [5,6,14–16,18]. However, for the most part these are asymptotic results that do not specify the particular constants involved, and leave gaps in the parameter possibilities where the results are silent. In particular, the empirical matchability demonstrations in this paper are not predictable from the previously known matchability asymptotics. Many of the known matchability results explicitly or implicitly involve edge correlation ϱ_e. The formulation of ϱ_h is new to this paper, and ϱ_T is also new to this paper. Thus we are now opening a fertile new avenue for proof-of-matchability results based on ϱ_h and ϱ_T, in the spirit of the existing results for ϱ_e and also for more powerful types of results.

Acknowledgments

The authors are grateful to The Maryland Advanced Research Computing Center for use of their supercomputer to conduct the computational experiments. An anonymous contributor made a very useful observation which led to streamlining the main result's proof. The referees' and editor's feedback and remarks greatly strengthened this article, and are very much appreciated. Our research was sponsored by the Air Force Research Laboratory and DARPA, under agreement numbers FA8750-18-2-0035 and FA8750-17-2-0112. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as representing official policies or endorsements, expressed or implied, of Air Force Research Laboratory, DARPA, or the U.S. Government.

Appendix A

We here provide some details about correlated Bernoulli random graphs. Notation here is as defined in the article.

Section A: For any $\{i,j\} \in \binom{[n]}{2}$ such that $0 < p_{i,j} < 1$, suppose that $\mathbb{1}(i \sim_G j)$ is a Bernoulli($p_{i,j}$) random variable and $\mathbb{1}(i \sim_H j)$ is a Bernoulli($p_{i,j}$) random variable, and suppose that the two random variables $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ have Pearson correlation coefficient ϱ_e; we derive the joint distribution of $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ as follows:

$$\begin{aligned}\varrho_e &= \frac{\mathrm{Cov}[\mathbb{1}(i \sim_G j), \mathbb{1}(i \sim_H j)]}{\sqrt{\mathrm{Var}[\mathbb{1}(i \sim_G j)]}\sqrt{\mathrm{Var}[\mathbb{1}(i \sim_H j)]}} \\ &= \frac{\mathbb{E}[\mathbb{1}(i \sim_G j)\mathbb{1}(i \sim_H j)] - \mathbb{E}[\mathbb{1}(i \sim_G j)] \cdot \mathbb{E}[\mathbb{1}(i \sim_H j)]}{\sqrt{p_{i,j}(1-p_{i,j})}\sqrt{p_{i,j}(1-p_{i,j})}} \\ &= \frac{\mathbb{P}[i \sim_G j \ \&\ i \sim_H j] - p_{i,j}^2}{p_{i,j}(1-p_{i,j})},\end{aligned}$$

from which we obtain $\mathbb{P}[i \sim_G j \ \&\ i \sim_H j] = p_{i,j}^2 + \varrho_e p_{i,j}(1-p_{i,j})$. Because $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ are each marginally Bernoulli($p_{i,j}$), we obtain that $\mathbb{P}[i \sim_G j \ \&\ i \not\sim_H j] = \mathbb{P}[i \not\sim_G j \ \&\ i \sim_H j] = p_{i,j} - \left(p_{i,j}^2 + \varrho_e p_{i,j}(1-p_{i,j})\right) = (1-\varrho_e)p_{i,j}(1-p_{i,j})$, and also that $\mathbb{P}[i \not\sim_G j \ \&\ i \not\sim_H j] = (1-p_{i,j}) - (1-\varrho_e)p_{i,j}(1-p_{i,j}) = (1-p_{i,j})^2 + \varrho_e p_{i,j}(1-p_{i,j})$.

Importantly, note that the joint distribution of $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ is uniquely determined by ϱ_e. Also note that $\mathbb{P}[\mathbb{1}(i \sim_G j) \neq \mathbb{1}(i \sim_H j)] = 2(1-\varrho_e)p_{i,j}(1-p_{i,j})$. Also note that, conditioned on $\mathbb{1}(i \sim_G j)$, the random variable Bernoulli($\varrho_e \cdot \mathbb{1}(i \sim_G j) + (1-\varrho_e) \cdot p_{i,j}$) results in the joint distribution above, which justifies the method in the article of sampling $\mathbb{1}(i \sim_G j)$ and $\mathbb{1}(i \sim_H j)$ with marginal Bernoulli($p_{i,j}$) distribution and Pearson correlation coefficient ϱ_e. □

Section B: We show that $\varrho_h \leq 1$, with equality holding if and only if, for all $\{i,j\} \in \binom{[n]}{2}$, it holds that $p_{i,j}$ is 0 or 1. Indeed,

$$\begin{aligned}&1 - \varrho_h \\ &= 1 - \frac{\sigma^2}{\mu(1-\mu)} \\ &= \frac{\mu(1-\mu) - \left(\frac{\sum_{\{i,j\} \in \binom{[n]}{2}} p_{i,j}^2}{\binom{n}{2}} - \mu^2\right)}{\mu(1-\mu)} \\ &= \frac{\sum_{\{i,j\} \in \binom{[n]}{2}} (p_{i,j} - p_{i,j}^2)}{\binom{n}{2}\mu(1-\mu)}\end{aligned}$$

is clearly nonnegative and equals 0 if and only if, for all $\{i,j\} \in \binom{[n]}{2}$ it holds that $p_{i,j} = p_{i,j}^2$, i.e. it holds that $p_{i,j}$ is 0 or 1. Thus $\varrho_h \leq 1$ with equality holding if and only if, for all $\{i,j\} \in \binom{[n]}{2}$, it holds that $p_{i,j}$ is 0 or 1. (Except, recall, the Bernoulli parameters are not all 0 and are not all 1, since ϱ_h would then not be defined.) □

References

[1] E. Airoldi, D. Blei, S.E. Fienberg, E. Xing, Mixed membership stochastic blockmodels, J. Mach. Learn. Res. 9 (2008) 1981–2014.
[2] A. Athreya, D. Fishkind, M. Tang, C. Priebe, Y. Park, J. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, Statistical inference on random dot product graphs: a survey, J. Mach. Learn. Res. 18 (2017) 8393–8484.
[3] P. Bickel, A. Chen, A nonparametric view of network models and newman–girvan and other modularities, Proc. Natl. Acad. Sci. 106 (2009) 21068–21073.
[4] D. Conte, P. Foggia, C. Sansone, M. Vento, Thirty years of graph matching in pattern recognition, Int. J. Pattern Recognit. Artif. Intell. 18:3 (2004) 265298.
[5] D. Cullina, N. Kiyavash, Improved achievability and converse bounds for erdos-renyi graph matching, ACM SIGMETRICS Perform. Eval. Rev. 44 (2016) 63–72.
[6] D. Cullina, N. Kiyavash, Exact alignment recovery for correlated erdos renyi graphs, arXiv:1711.06783 (2017).
[7] J. Ding, Z. Ma, Y. Wu, J. Xu, Efficient random graph matching via degree profiles, arXiv:1811.07821 (2018).
[8] D. Fishkind, S. Adali, H. Patsolic, L. Meng, D. Singh, V. Lyzinski, C. Priebe, Seeded graph matching, Pattern Recognit. 87 (2019) 203–215.
[9] P. Foggia, G. Percannella, M. Vento, Graph matching and learning in pattern recognition in the last 10 years, Int. J. Pattern Recognit Artif. Intell. 28:1 (2014).
[10] P. Hoff, A. Raftery, M. Handcock, Latent space approaches to social network analysis, J. Am. Stat. Assoc. 97 (2002) 1090–1098.
[11] P. Holland, K. Laskey, S. Leinhardt, Stochastic blockmodels: first steps, Soc. Networks 5 (1983) 109–137.
[12] J. Lei, A goodness-of-fit test for stochastic block models, Ann. Stat. 44 (2016) 401–424.
[13] K. Levin, A. Athreya, M. Tang, V. Lyzinski, Y. Park, C. Priebe, A central limit theorem for an omnibus embedding of random dot product graphs, arXiv:1705.09355 (2017).
[14] V. Lyzinski, Information recovery in shuffled graphs via graph matching, IEEE Trans. Inf. Theory 64:5 (2018) 3254–3273.
[15] V. Lyzinski, D. Fishkind, M. Fiori, J. Vogelstein, C. Priebe, G. Sapiro, Graph matching: relax at your own risk, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 60–73.
[16] V. Lyzinski, D. Fishkind, C. Priebe, Seeded graph matching for correlated Erdos–Renyi graphs, J. Mach. Learn. Res. 15 (2014) 3693–3720.
[17] V. Lyzinski, M. Tang, A. Athreya, Y. Park, C. Priebe, Community detection and classification in hierarchical stochastic blockmodels, IEEE Trans. Network Sci. Eng. 4 (2017) 13–26.
[18] E. Onaran, S. Garg, E. Erkip, Optimal de-anonymization in random graphs with community structure, 2016 IEEE 37th Sarnoff Symposium, 2016.
[19] H. Patsolic, Y. Park, V. Lyzinski, C. Priebe, Vertex nomination via local neighborhood matching, arXiv:1705.00674 (2017).
[20] C. Priebe, Y. Park, M. Tang, A. Athreya, V. Lyzinski, J. Vogelstein, Y. Qin, B. Cocanougher, K. Eichler, M. Zlatic, A. Cardona, Semiparametric spectral modeling of the drosophila connectome, arXiv:1705.03297 (2017).
[21] K. Rohe, S. Chatterjee, B. Yu, Spectral clustering and the high-dimensional stochastic blockmodel, Ann. Stat. 39 (2011) 1878–1915.

[22] D. Sussman, M. Tang, C. Priebe, Consistent latent position estimation and vertex classification for random dot product graphs, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 48–57.
[23] M. Tang, A. Athreya, D. Sussman, V. Lyzinski, Y. Park, C. Priebe, A semiparametric two-sample hypothesis testing problem for random dot product graphs, J. Comput. Graph. Stat. 26 (2017) 344–354.
[24] M. Tang, A. Athreya, D. Sussman, V. Lyzinski, C. Priebe, A nonparametric two-sample hypothesis testing problem for random dot product graphs, Bernoulli 23 (2017) 1599–1630.
[25] J. Vogelstein, J. Conroy, V. Lyzinski, L. Podrazik, S. Kratzer, E. Harley, D. Fishkind, R. Vogelstein, C. Priebe, Fast approximate quadratic programming for graph matching, PLoS ONE 10:4 (2015) e0121002.
[26] K. Xu, A. Hero, Dynamic stochastic blockmodels for time-evolving social networks, IEEE J. Sel. Top Signal Process. 8 (2014) 552–562.
[27] S. Young, E. Scheinerman, Random dot product graph models for social networks, in: Proceedings of the 5th International Conference on Algorithms and Models for the Web-Graph, 2007, pp. 138–149.