Computational Statistics and Data Analysis 54 (2010) 1777–1790

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Statistical inference on attributed random graphs: Fusion of graph features and content

John Grothendieck [a], Carey E. Priebe [b,*], Allen L. Gorin [c]

[a] *BBN technologies, United States*
[b] *Johns Hopkins University, United States*
[c] *US Department of Defense, United States*

ARTICLE INFO

Article history:
Received 8 June 2009
Received in revised form 10 January 2010
Accepted 10 January 2010
Available online 2 February 2010

Keywords:
Information fusion
Statistical inference
Random graphs

ABSTRACT

Many problems can be cast as statistical inference on an attributed random graph. Our motivation is change detection in communication graphs. We prove that tests based on a fusion of graph-derived and content-derived metadata can be more powerful than those based on graph or content features alone. For some basic attributed random graph models, we derive fusion tests from the likelihood ratio. We describe the regions in parameter space where the fusion improves power, using both numeric results from selected small examples and analytic results on asymptotically large graphs.

1. Introduction

There are many problems that can be cast as statistical inference on an attributed random graph. An example commonly encountered in language processing is a communication graph, in which vertices represent entities and edges the messages between them. Thus edges and vertices are both complex objects with many potential attributes of interest. Graph properties such as degree and adjacency provide a context for inference; see Grothendieck et al. (2009) on Switchboard and Priebe et al. (2005) on the Enron corpus. Similar attributed graphs emerge from social network analysis (Leenders, 1995), internet traffic (Sen et al., 2004), and in applications of entity-relation extraction (Doddington et al., 2004).

A great deal is known about random graphs (Bollobas, 2001). Certain attributed graphs (e.g. k-colorings) are also familiar. Yet the intersection – random graphs with attributes – has a sparse literature. This is a regrettable omission; certain tasks such as understanding natural language benefit from context beyond the content itself, and a communication graph is a natural model for capturing elements of context. Our goal is to move beyond intuition to an analytic understanding of how and under what circumstances the additional context provided by an attributed graph aids statistical inference.

This work considers the inference problem of detecting departures from some ordinary state, which shall serve as our null model of an attributed graph. The motivating application is change detection — the null model is constructed from past observations in some time window, and present observations may or may not be a good fit with that past model. We consider several basic models of edge-attributed random graphs, building upon the notion of an Erdõs–Rényi (ER) random graph (Gilbert, 1959). Our null models are homogeneous random graphs with one edge feature. Models for the alternative case include a homogeneous random graph with different parameters, or a heterogeneous graph with an anomalous block of vertices. Some natural generalizations of these structures suggest themselves.

* Corresponding address: Johns Hopkins University, Whiting School of Engineering, MD 21218-2682 Baltimore, United States. Tel.: +1 410 516 7200; fax: +1 410 516 7459.
E-mail address: cep@jhu.edu (C.E. Priebe).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.017

In the case of a communication graph such as the Enron email corpus, message attributes can be divided into classes Ξ_C, Ξ_E, and Ξ_G for content, externals, and graph features, respectively. Less briefly, these are: features extracted from content (e.g. topic classification of the message body), features external to the message body (e.g. header fields such as "Date":), and features that depend on graph context beyond the message itself (e.g. an identical message was just received by everyone in the company). For our inference problem, we demonstrate that a test using both graph and content features can be more powerful than testing either alone. Different tests are best at different points in the set of alternatives; we prove that, under certain conditions, fusion is most helpful in the region where the individual features have similar power.

The remainder of this paper is organized as follows. In Section 2, we analyze the simple case of detecting a homogeneous alternative. Section 3 considers the case of detecting an anomalous block of vertices, derives the asymptotic form of the likelihood ratio test, and presents power improvement results on simulation data. In Section 4, we extend our models to more general families of attributed random graphs and consider the impact on asymptotic hypothesis testing. Conclusions are presented in Section 5.

2. Attributed random graphs: Global alternative

In this section we consider the simplest case of interest. Here the hypothesis test is whether the stochastic parameters of a basic attributed graph model take the values under the null hypothesis. We present the random graph model and derive the likelihood ratio (LR) test. This provides a proof that fusion tests can be more powerful than tests based on the graph or content features alone; several other results of analytic interest are immediate consequences.

2.1. ER_c model

Consider an attributed graph $G = (V, A)$ where $V = \{1, \ldots, n\}$. For simplicity, edges A possess one categorical attribute, so that each $e \in A$ is given by a triple $(u, v, l) \in V \times V \times \mathcal{L}$. In our conception, $(u, v, l) \in A$ is a *communication* between vertex u and vertex v. The pair (u, v) represents the *externals* of the communication which has *content* labeled by *topic* $\ell \in \mathcal{L}$, the *topic set*. The *graph features* are extracted from the overall collection of externals. We consider $|\mathcal{L}| = k < \infty$, and *simple* graphs G — undirected, without graph loops, and having at most one edge between any two vertices.

We first consider the simplest random graph models that admit hypothesis testing based upon both graph features and content. For integer $n \geq 2$ and $p, c \in [0, 1]$, let $\mathrm{ER}_c(n, p, c)$ denote the random graph model on vertices $V = \{1, \ldots, n\}$ such that each unordered pair of distinct vertices is joined by an edge according to independent Bernoulli(p) random variables, as in the classic ER random graph. In addition, each edge has associated with it topic $l \in \mathcal{L} = \{0, 1\}$ according to independent Bernoulli(c) random variables. The notation ER_c is meant to suggest an Erdõs–Rényi graph "with content" or "with color" on the edges. Notice that consideration of unattributed graph structure alone gives rise to an $\mathrm{ER}(n, p)$ random graph model, identified with all edges and denoted A, and that consideration of content alone gives rise to an $\mathrm{ER}(n, pc)$ random graph model, identified with all edges labeled with topic $l = 1$ and denoted C.

Our null hypothesis shall be an ER_c graph. We investigate an alternative scenario involving another ER_c graph with distinct stochastic generation of both graph features and content. Then one may consider statistical power based on one aspect alone, or using their joint distribution — a combination or *fusion* of features.

For integer $n \geq 2$ and p_0, c_0 in the interval $[0, 1]$, let the null hypothesis H_0 be an $\mathrm{ER}_c(n, p_0, c_0)$ graph. This contrasts with alternative H_A an $\mathrm{ER}_c(n, p_A, c_A)$ graph, for $p_A, c_A \in [0, 1]$ with $p_A \geq p_0$, $c_A \geq c_0$. We denote the set of alternative parameters $\Theta_A = [p_0, 1] \times [c_0, 1] \setminus \{(p_0, c_0)\}$, where $\setminus$ denotes set subtraction. These basic alternative models provide analytic insight before we consider the non-homogeneous models of Section 3.

2.2. Log-likelihood ratio for fusion

Consider the test statistics $T_1 = |A|$ based upon graph features only, $T_2 = |C|$ based upon content only, and $s(T_1, T_2)$ some fusion of both (to be defined later). These are random variables, defined as functions of the random graph model. What form do tests of a simple null versus simple alternative take?

An ER graph with n vertices has $\binom{n}{2}$ potential edges, each of which exist according to independent Bernoulli random variables. Thus, under the $\mathrm{ER}_c(n, p_0, c_0)$ null hypothesis, the statistics are distributed as binomial random variables:

$$T_1 \sim_{H_0} \mathrm{Bin}\left(\binom{n}{2}, p_0\right)$$

$$T_2 \sim_{H_0} \mathrm{Bin}\left(\binom{n}{2}, c_0 p_0\right).$$

Under the $\mathrm{ER}_c(n, p_A, c_A)$ alternative

$$T_1 \sim_{H_A} \mathrm{Bin}\left(\binom{n}{2}, p_A\right)$$

$$T_2 \sim_{H_A} \mathrm{Bin}\left(\binom{n}{2}, c_A p_A\right).$$

T_1 and T_2 are dependent, since only edges that exist are given topics. In particular, using lower-case t_1 and t_2 to represent observed values of the corresponding random variables

$$T_2 \mid (T_1 = t_1) \sim_{H_0} \text{Bin}(t_1, c_0)$$
$$\sim_{H_A} \text{Bin}(t_1, c_A).$$

Thus we have likelihood functions

$$f_0(t_1, t_2) = \binom{\binom{n}{2}}{t_1} p_0^{t_1}(1-p_0)^{\binom{n}{2}-t_1} \binom{t_1}{t_2} c_0^{t_2}(1-c_0)^{t_1-t_2}$$
$$f_A(t_1, t_2) = \binom{\binom{n}{2}}{t_1} p_A^{t_1}(1-p_A)^{\binom{n}{2}-t_1} \binom{t_1}{t_2} c_A^{t_2}(1-c_A)^{t_1-t_2}.$$

The Neyman–Pearson Lemma states that the most powerful (MP) test of a simple null versus simple alternative hypothesis is the (log-)likelihood ratio (LR) test. For a particular (p_A, c_A), log-likelihood ratio $\lambda = \log f_A/f_0$ is

$$\begin{aligned}\lambda_{(p_A,c_A)}(t_1, t_2) &= t_1 \log \frac{p_A}{p_0} + \left(\binom{n}{2} - t_1\right) \log \frac{1-p_A}{1-p_0} + t_2 \log \frac{c_A}{c_0} + (t_1 - t_2) \log \frac{1-c_A}{1-c_0} \\ &= t_1 \left(\log \frac{p_A}{p_0} + \log \frac{1-p_0}{1-p_A} - \log \frac{1-c_0}{1-c_A}\right) + t_2 \left(\log \frac{c_A}{c_0} + \log \frac{1-c_0}{1-c_A}\right) + \binom{n}{2} \log \frac{1-p_A}{1-p_0} \\ &= \gamma_1 t_1 + \gamma_2 t_2 + \gamma_3. \end{aligned} \tag{1}$$

This test statistic $\lambda_{(p_A,c_A)}$ is a linear combination of observed (t_1, t_2), and the MP test rejects the null for values of $\lambda_{(p_A,c_A)}$ above some threshold.

Theorem 2.1. *For almost any $(p_A, c_A) \in \Theta_A$, there exists a test using (T_1, T_2) that is more powerful at that point than a test based on T_1 or T_2 alone.*

Proof. By the Neyman–Pearson Lemma. The likelihood ratio above provides such a test, save in cases where $\gamma_1 = 0$ or $\gamma_2 = 0$. ■

Different values (p_A, c_A) can lead to different $\lambda_{(p_A,c_A)}$. Over a compound (more than one-point) alternative hypothesis, there may be no uniformly most powerful (UMP) test.

Corollary 2.2. *No UMP test using (T_1, T_2) exists for $H_0 : \text{ER}_c(n, p_0, c_0)$ versus $H_A : \text{ER}_c(n, p_A, c_A)$ over all Θ_A.*

Proof. The critical value of a likelihood ratio is determined by the desired test level (probability of rejecting the null when it is true). It is well known that a linear transform of any test statistic leads to a test of equivalent power. Thus adding a constant to γ_3 in the LR statistic above leads to an equivalent test, simply with a different critical value. Similarly, scalar multiplication leads to a scalar change in the critical value and an equivalent test. Hence all $\lambda_{(p_A,c_A)}$ with the same ratio γ_1/γ_2 (i.e. same projective coefficient) are equivalent test statistics. However, the MP tests on different points in Θ_A have different values of this ratio, defining non-equivalent statistics of (T_1, T_2). No UMP test exists on all of Θ_A. ■

Corollary 2.3. *There exist curves Λ_r in Θ_A on which a UMP test using (T_1, T_2) exists.*

Proof. Those (p_A, c_A) with the same ratio $r = \gamma_1/\gamma_2$ define curves in Θ_A:

$$\begin{aligned}\Lambda_r &= \left\{(p_A, c_A) : \frac{c_A^r(1-c_0)^{r+1}}{c_0^r(1-c_A)^{r+1}} = \frac{p_A(1-p_0)}{p_0(1-p_A)}\right\} \\ &= \left\{(p_A, c_A) : p_A = \left[1 + \frac{(1-p_0)c_0^r(1-c_A)^{r+1}}{p_0 c_A^r(1-c_0)^{r+1}}\right]^{-1}\right\}.\end{aligned}$$

On each Λ_r, the individual $\lambda_{(p_A,c_A)}$ define equivalent tests. This test is MP for each point in Λ_r, hence UMP if attention is restricted to those points. ■

Note the curve on which the coefficient for T_1 is zero:

$$1 - c_A = \frac{p_0(1-p_A)(1-c_0)}{p_A(1-p_0)}$$
$$c_A = \frac{c_0 p_0(1-p_A) + p_A - p_0}{p_A(1-p_0)}.$$

Neglecting the degenerate cases with zero Bernoulli probabilities, this always returns a valid $c_A \in [c_0, 1]$. Rearranging terms,

$$\Delta(cp) = K\Delta(p)$$
$$(c_A p_A - c_0 p_0) = \frac{1 - c_0 p_0}{1 - p_0}(p_A - p_0)$$

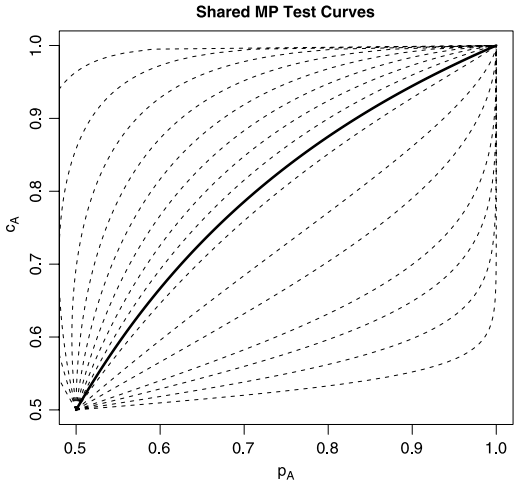

Fig. 1. Λ_r for selected values of projective coordinate r, with $p_0 = c_0 = 0.5$. The most powerful test uses both T_1 and T_2 for any point not on solid curve Λ_0.

defines a line in the coordinate system based upon Bernoulli parameter differences. For c_A below Λ_0, the best test puts a positive coefficient on T_1; above it, the somewhat non-intuitive result that larger T_1 implies *less* relative likelihood for the alternative hypothesis. Fig. 1 shows these curves. This also demonstrates that fusion cannot always improve power. The UMP test over Λ_0 rejects for T_2 above some threshold, independent of T_1. While Λ_0 is of measure zero (given a Lebesgue prior on parameter space), its existence indicates that a fusion test will not have improved power over all possible alternatives.

In practice, we are unlikely to have a simple alternative hypothesis, or to conveniently know that the alternative lies in some Λ_r. The standard extension of the simple-versus-simple LR test addresses this by comparing the members of the set of nulls/alternatives with the greatest likelihood. Maximum likelihood (ML) parameter estimates are not quite trivial to calculate; unrestricted parameters lead to ML estimates

$$\hat{p}_A = \frac{t_1}{\binom{n}{2}}, \qquad \hat{c}_A = \frac{t_2}{t_1}.$$

However, there are special cases to consider where the ML parameters lie on the boundary of Θ_A. Replacing fixed p_A, c_A in Eq. (1) with ML estimates provides the general LR test, with λ no longer dependent on unknown (p_A, c_A).

2.3. Alternative coordinates

The "natural" coordinates for the parameter space are not obvious, even in this simple example. Our interest in attributed random graphs and the combination of graphical and content-derived metadata motivates parameters (p_A, c_A). Mathematically, the parameters $(p_A, c_A p_A)$ are natural for binomial random variables T_1 and T_2 respectively, while their joint distribution might be more easily modeled via a trinomial random variable (i.e. edges are absent, topic 1, or topic 2). See Fig. 2. While this work primarily uses (p_A, c_A) for internal consistency, the authors do invoke different parameterizations where these provide insight.

3. Attributed random graphs: Block alternative

3.1. Simple blockmodel

We have considered random graph models under some form of *homogeneity* — all pairs of vertices have the same stochastic properties in terms of communication (graph structure and content). In this section, we consider a more complex alternative model. The alternative *stochastic blockmodel* hypothesis is that the vertex set V is partitioned into subsets such that the stochastic properties of communications between a pair of vertices depend only on the subset membership of each (see for example Airoldi et al., 2008). Thus, when vertices are ordered by subset, an $n \times n$ matrix of a stochastic feature such as edge probabilities P_{ij} is a block matrix. Blockmodels better fit many communication graphs — people interact with friends or coworkers more often than with random strangers.

The simplest nontrivial such model has two subsets V^0 and V^A such that the stochastic properties of any pair of vertices in V^0, or of a pair of vertices one from V^0 and one from V^A, is the same as under the null while the stochastic properties of any pair of vertices in V^A differs from that under the null. The standard blockmodel further allows different stochastic properties for edges connecting V^0 and V^A; the simpler model is motivated by the smaller difference from null, with particular application to security problems. That is, a group of actors involved in nefarious activity might present as "typical" a profile towards those outside the group as they could contrive, in order to avoid notice.

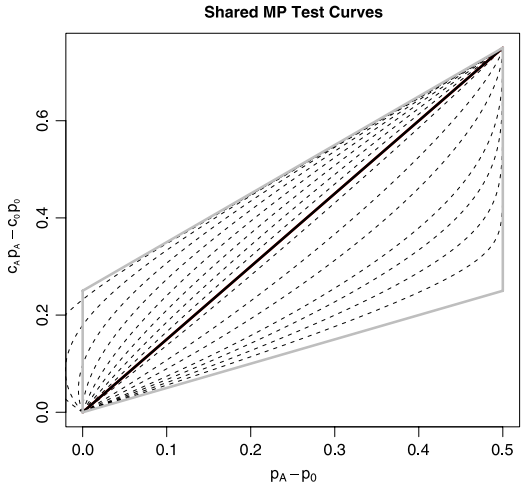

Fig. 2. Λ_r for selected values of r, with $p_0 = c_0 = 0.5$, using $(p_A, c_A p_A)$ coordinates.

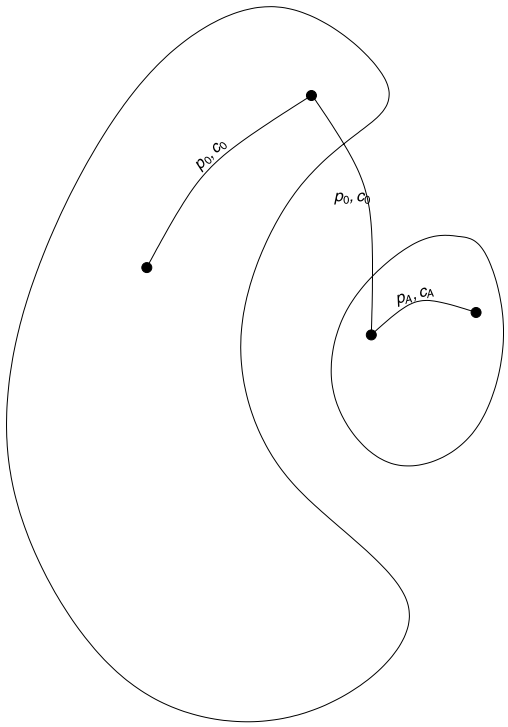

Fig. 3. Sketch of alternative hypothesis $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$. As with ER_c, subscript-c denotes "with content". The small "egg" represents the m vertices with anomalous communication activity amongst them, via parameters p_A and c_A. The large "kidney" represents the remaining $n - m$ vertices, and communication activity between any pair of these vertices or between one of these kidney vertices and one of the egg vertices behaves as in the $\mathrm{ER}_c(n, p_0, c_0)$ null model.

Alternative H_A is the graph model, denoted $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$, in which there is a subset V^A of V with cardinality $|V^A| = m \in \{2, \ldots, n\}$. (Note that $|V^A| \geq 2$ in order for the alternative to differ from the null.) Each unordered pair of distinct vertices in V^A has an edge according to independent Bernoulli (p_A) random variables and each edge has associated topic $l \in \{0, 1\}$ according to independent Bernoulli (c_A) random variables, and all other pairs of vertices have edge probability p_0 and topic probability c_0 as in $\mathrm{ER}_c(n, p_0, c_0)$. Thus the $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$ alternative can be conceived as a communication graph in which there is a subset of vertices with increased probability of activity and that activity involves an increased probability of topic $l = 1$. Fig. 3 provides a sketch of $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$. We assume $2 \leq m \leq n - 2$ so that both V^A and $V \setminus V^A$ contain at least one pair of vertices.

3.2. Finite sample results: Analysis and theorem

Under the $\mathrm{ER}_c(n, p_0, c_0)$ null hypothesis, we have

$$T_1 \sim_{H_0} \mathrm{Bin}\left(\binom{n}{2}, p_0\right) \tag{2}$$

while under the $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$ alternative, T_1 is the sum of two *independent* binomial random variables

$$T_1 \sim_{H_A} \mathrm{Bin}\left(\binom{n}{2} - \binom{m}{2}, p_0\right) +_{ind} \mathrm{Bin}\left(\binom{m}{2}, p_A\right). \tag{3}$$

Similarly,

$$T_2 \sim_{H_0} \mathrm{Bin}\left(\binom{n}{2}, c_0 p_0\right) \tag{4}$$

and

$$T_2 \sim_{H_A} \mathrm{Bin}\left(\binom{n}{2} - \binom{m}{2}, c_0 p_0\right) +_{ind} \mathrm{Bin}\left(\binom{m}{2}, c_A p_A\right). \tag{5}$$

Define T_1^A (resp. T_2^A) to be the total counts in T_1 (resp. T_2) among the $\binom{m}{2}$ possible edges in V^A. Likewise define T_1^0 and T_2^0 to be the counts from the $\binom{n}{2} - \binom{m}{2}$ edges with the same stochastic properties as under the null hypothesis. Note that $T_1 = T_1^0 + T_1^A$ and $T_2 = T_2^0 + T_2^A$.

T_2 is dependent on T_1. As well as the range of possible values of T_2, T_1 also impacts the "average" content parameter by providing information about the relative size of T_1^A. Let $a(T_1)$ be the expected proportion of existing edges among V^A. In particular, there is conditional distribution

$$\begin{aligned} T_2 \mid (T_1 = t_1) &\sim_{H_0} \mathrm{Bin}(t_1, c_0) \\ &\sim_{H_A} \mathrm{Bin}(t_1, \tilde{c}) \end{aligned}$$

where

$$\tilde{c} = a(t_1)c_A + (1 - a(t_1))c_0, \quad \text{and} \quad a(t_1) = \frac{1}{t_1}\mathbb{E}_A\left(T_1^A \mid T_1 = t_1\right).$$

In principle we can directly calculate critical values and test powers using these null and alternative distributions, simply by summing over the appropriate binomial count probabilities. For large n, this is not tractable, but does allow the following result.

Theorem 3.1. *A test based on fusion of graph features and content can be more powerful than a test based on either one alone.*

Proof. By construction, we provide an example wherein the result holds.

Define H_0, H_A, T_1, and T_2 as in the rest of Section 3.

Define $T = s(T_1, T_2)$: test statistic based on fusion of graph features and content.

Let β_1 be the power of a test based upon T_1, β_2 the power of a test based upon T_2.

We demonstrate that $\beta > \max\{\beta_1, \beta_2\}$ for appropriate choices of $n, p_0, c_0, m, p_A, c_A, \alpha$, and T.

Consider $T = T_1 + T_2$. Under the alternative, $(T_1^0 + T_2^0)$ is independent of $(T_1^A + T_2^A)$. Thus we need only identify the distributions of the independent conditional random variables $Z^0 = T_2^0|_{T_1^0 = t_1^0}$, which is Binomial (t_1^0, c_0), and $Z^A = T_2^A|_{T_1^A = t_1^A}$, which is Binomial (t_1^A, c_A). From these distributions we can directly calculate the critical value and the power value for the test based on T.

It remains to calculate rejection probabilities and compare tests at the same level of falsely rejecting the null. Calculating alternative power is then simply a matter of summing binomial probabilities over all combinations of kidney/egg edges that would exceed the critical threshold. (Decisions may be randomized at threshold counts T_1 or T_2 to achieve any desired test level α.) Fig. 4 presents results for one case showing $\beta > \max\{\beta_1, \beta_2\}$ on a region in Θ_A. ■

3.3. Asymptotic distributions

Analytic structure might be more apparent (and certainly more tractable) using the asymptotic distributions as number of vertices $n \to \infty$. For notational convenience, we define

$$x_1 = \frac{T_1 - \binom{n}{2} p_0}{\sqrt{\binom{n}{2} p_0(1 - p_0)}}, \qquad \mu_1 = \frac{\binom{m}{2}(p_A - p_0)}{\sqrt{\binom{n}{2} p_0(1 - p_0)}}$$

$$x_2 = \frac{T_2 - \binom{n}{2} c_0 p_0}{\sqrt{\binom{n}{2} c_0 p_0(1 - c_0 p_0)}}, \qquad \mu_2 = \frac{\binom{m}{2}(c_A p_A - c_0 p_0)}{\sqrt{\binom{n}{2} c_0 p_0(1 - c_0 p_0)}}.$$

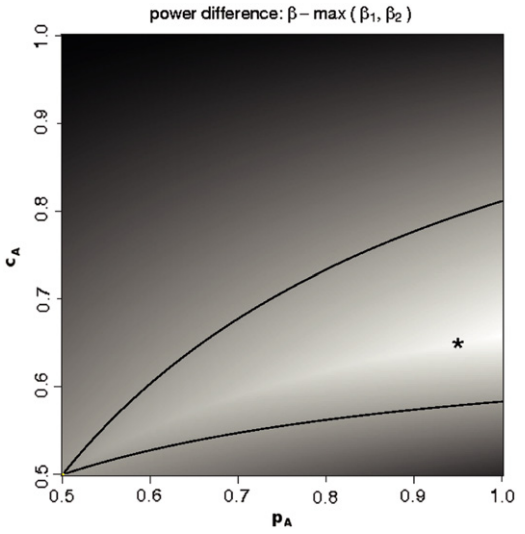

Fig. 4. Figure demonstrating the superiority of fusion of graph features and content. This result is for $s_{0.5}(x, y) = \frac{x+y}{2}$ with $n = 5$, $p_0 = c_0 = 0.5$, $\alpha = 0.05$, with fixed $m = 2$ and (p_A, c_A) varying throughout alternative region Θ_A. Lighter shades indicate larger values of $\beta - \max\{\beta_1, \beta_2\}$, solid curves indicates values of (p_A, c_A) for which $\beta - \max\{\beta_1, \beta_2\} = 0$, and the region within the solid curves is the collection of values (p_A, c_A) for which $\beta - \max\{\beta_1, \beta_2\} > 0$. e.g., at $(p_A = 0.95, c_A = 0.65)$ – the star – we obtain $\beta = 0.083 > \max\{\beta_1 = 0.079, \beta_2 = 0.079\}$.

By the Central Limit Theorem, as $n \to \infty$,

$$x_1 \overset{\mathcal{D}}{\underset{\to H_0}{}} Z(0, 1)$$

$$\overset{\mathcal{D}}{\underset{\to H_A}{}} Z\left(\mu_1, \frac{\binom{m}{2} p_A(1-p_A) + \left(\binom{n}{2} - \binom{m}{2}\right) p_0(1-p_0)}{\binom{n}{2} p_0(1-p_0)}\right)$$

$$x_2 \overset{\mathcal{D}}{\underset{\to H_0}{}} Z(0, 1)$$

$$\overset{\mathcal{D}}{\underset{\to H_A}{}} Z\left(\mu_2, \frac{\binom{m}{2} c_A p_A(1-c_A p_A) + \left(\binom{n}{2} - \binom{m}{2}\right) c_0 p_0(1-c_0 p_0)}{\binom{n}{2} c_0 p_0(1-c_0 p_0)}\right).$$

Note the consequence that power goes to α in the limit if m is of order less than $n^{0.5}$, with V^A too small to impact global statistics T_1 and T_2. Likewise power goes to one if m grows faster than $n^{0.5}$, with V^A large enough to detect with no possible improvement from fusion. Asymptotically the interesting case is where m grows exactly as $n^{0.5}$. Henceforth we assume such an order relationship, yielding

$$x_1 \overset{\mathcal{D}}{\underset{\to H_A}{}} Z(\mu_1, 1), \qquad x_2 \overset{\mathcal{D}}{\underset{\to H_A}{}} Z(\mu_2, 1).$$

Another consequence is to suggest more natural coordinates for Fig. 4. We could base these upon the normalized statistics for tests based on T_1 and T_2 alone. Thus each value along the x-axis (or y-axis) would correspond to a particular β_1 (or β_2), as in Figs. 5 and 6.

T_1, T_2, and the T from the proof of Theorem 3.1 are instances of more general

$$s_\gamma(T_1, T_2) = \gamma T_1 + (1-\gamma) T_2,$$

$$x_\gamma = \frac{s_\gamma(T_1, T_2) - \binom{n}{2}(\gamma p_0 + (1-\gamma) c_0 p_0)}{\sqrt{\binom{n}{2} p_0 \left(\gamma^2(1-p_0) + 2\gamma(1-\gamma) c_0(1-p_0) + (1-\gamma)^2 c_0(1-c_0 p_0)\right)}}.$$

Then there is asymptotically normal

$$x_\gamma \overset{\mathcal{D}}{\underset{\to H_0}{}} Z(0, 1)$$

$$\overset{\mathcal{D}}{\underset{\to H_A}{}} Z\left(\frac{\binom{m}{2}(\gamma(p_A - p_0) + (1-\gamma)(c_A p_A - c_0 p_0))}{\sqrt{\binom{n}{2} p_0 \left(\gamma^2(1-p_0) + 2\gamma(1-\gamma) c_0(1-p_0) + (1-\gamma)^2 c_0(1-c_0 p_0)\right)}}, 1\right).$$

For fixed (p_0, c_0, p_A, c_A), the alternative mean has a critical value w.r.t.

$$\gamma = \frac{c_0(1-c_0 p_0)(p_A - p_0) - (1-p_0)(c_A p_A - c_0 p_0)}{(1-c_0)\left(c_0 p_0(p_A - p_0) + (1-p_0)(c_A p_A - c_0 p_0)\right)}.$$

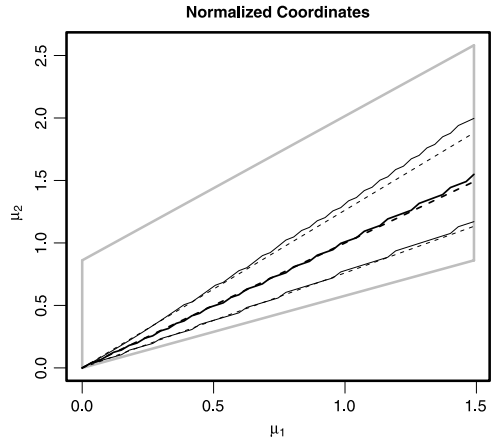

Fig. 5. Here axes are alternative expected values of normalized test statistics for T_1 and T_2 alone, μ_1 and μ_2. $s_{0.5}(x, y) = \frac{x+y}{2}$ and $n = 10$, $p_0 = c_0 = 0.5$, $\alpha = 0.05$, and $m = 5$. The image of Θ_A lies within the dashed lines. Solid black lines are where $\beta - \max\{\beta_1, \beta_2\} = 0$. The solid gray line shows the best observed power improvement for fixed p_A. Dotted lines are asymptotic solutions to contrast with the empirical power results.

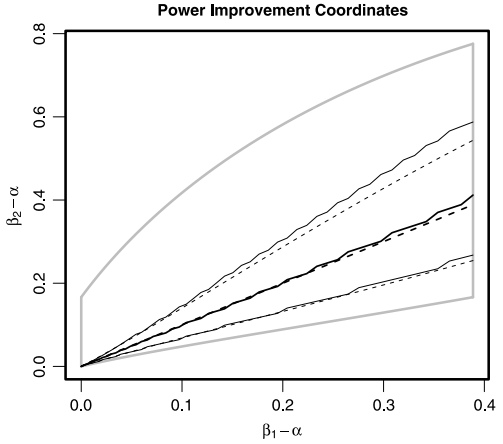

Fig. 6. Empirical power results as in Fig. 5, except here axes are alternative rejection probability increases $\beta_i - \alpha$ of the UMP level α hypothesis tests based on T_1 and T_2.

This defines the MP member of the family of fusion statistics $s_\gamma(T_1, T_2)$ with respect to detecting that particular $(p_A, c_A) \in \Theta_A$.

3.4. Asymptotic LLR

Direct calculation of T_1^A proportion $a(t_1)$ from the binomial sum of H_A becomes expensive — polynomial in $\binom{n}{2}$. Asymptotically T_1 and T_2 are Gaussian with known covariance, with a simple formula for conditional expectation:

$$a(t_1) \to \frac{1}{t_1}\left(\binom{m}{2} p_A + \frac{\binom{m}{2} p_A(1-p_A)\left[t_1 - \binom{n}{2} p_0 - \binom{m}{2}(p_A - p_0)\right]}{\binom{m}{2} p_A(1-p_A) + (\binom{n}{2} - \binom{m}{2}) p_0(1-p_0)}\right)$$
$$\to \frac{\binom{m}{2} p_A}{t_1} \to \frac{\binom{m}{2} p_A}{\binom{n}{2} p_0}. \tag{6}$$

Given our order assumptions $a(T_1) = O(n^{-1})$.

Consider LR statistic λ for simple null $\mathrm{ER}_c(n, p_0, c_0)$ against simple alternative $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$ using both T_1 and T_2. Conditionally, $T_2|T_1$ can be normalized via

$$x_{2|1} = \frac{T_2 - c_0 t_1}{\sqrt{t_1 c_0(1-c_0)}}, \qquad \mu_{2|1} = \frac{t_1 a(t_1)(c_A - c_0)}{\sqrt{t_1 c_0(1-c_0)}}.$$

This leads to

$$\begin{aligned}\lambda(T_1 = t_1, T_2 = t_2) &= \left[\mu_1 x_1 - \frac{\mu_1^2}{2}\right] + \left[\mu_{2|1} x_{2|1} - \frac{\mu_{2|1}^2}{2}\right] \\ &= t_1 \left[\frac{\binom{m}{2}(p_A - p_0)}{\binom{n}{2} p_0(1-p_0)} - \frac{a(t_1)(c_A - c_0)}{(1-c_0)} - \frac{a(t_1)^2 (c_A - c_0)^2}{2c_0(1-c_0)}\right] \\ &\quad + t_2 \frac{a(t_1)(c_A - c_0)}{c_0(1-c_0)} - \frac{\binom{m}{2}(p_A - p_0)}{1-p_0}\left[1 + \frac{\binom{m}{2}(p_A - p_0)}{2\binom{n}{2} p_0}\right]. \end{aligned} \tag{7}$$

Consider the coefficient on T_1:

$$\begin{aligned}\frac{\binom{m}{2}(p_A - p_0)}{\binom{n}{2} p_0(1-p_0)} - \frac{a(t_1)(c_A - c_0)}{(1-c_0)} - \frac{a(t_1)^2 (c_A - c_0)^2}{2c_0(1-c_0)} &\to \frac{\binom{m}{2}(p_A - p_0)}{\binom{n}{2} p_0(1-p_0)} - \frac{a(t_1)(c_A - c_0)}{(1-c_0)} \\ &\approx \frac{\binom{m}{2}(p_A - p_0)}{\binom{n}{2} p_0(1-p_0)} - \frac{\binom{m}{2} p_A (c_A - c_0)}{\binom{n}{2} p_0(1-c_0)}.\end{aligned}$$

Setting this equal to zero:

$$c_A p_A - c_0 p_0 \approx \left(c_0 + \frac{1-c_0}{1-p_0}\right)(p_A - p_0).$$

As in the global alternative case, there is in fact such $c_A \in (c_0, 1)$ for each $p_A \in (p_0, 1)$.

Theorem 3.2. *Asymptotically as $n \to \infty$, for almost any $(p_A, c_A) \in \Theta_A$, there exists a test using (T_1, T_2) that is more powerful at that point than a test based on T_1 or T_2 alone. No UMP test exists for $H_0 : \mathrm{ER}_c(n, p_0, c_0)$ versus $H_A : \mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$ over all of Θ_A. There exist curves in Θ_A on which a UMP test using (T_1, T_2) exists.*

Proof. Essentially the same arguments as for $H_A : \mathrm{ER}_c(n, p_A, c_A)$. ■

Note that each $\lambda_{(p_A, c_A)}$ test, as a linear combination of T_1 and T_2, is equivalent to a member of the family s_γ considered earlier. Each $\lambda_{(p_A, c_A)}$ is optimal on some subset of Θ_A. Recall tests $s_\gamma = \gamma T_1 + (1-\gamma) T_2$. We now provide some characterization of where these tests improve power over T_1 or T_2 alone.

Theorem 3.3. *Let the power of a test based on T_1 alone be β_1, and on T_2 alone be β_2. A test $s_\gamma(T_1, T_2)$ with $\gamma \in (0, 1)$ has greater power than $\max\{\beta_1, \beta_2\}$ in some region of Θ_A. Let β_γ be the power of the test rejecting for large s_γ. Power improvement $\beta_\gamma - \max\{\beta_1, \beta_2\}$ is greatest on the curve $\{(p_A, c_A) : \mu_1 = \mu_2 \subset \Theta_A\}$.*

Proof. We abuse notation and denote the normalized mean value of s_γ under the alternative by μ_γ. s_γ does as well as the test using T_1 alone when $\mu_1 = \mu_\gamma$, and as well as the test using T_2 alone when $\mu_2 = \mu_\gamma$. This describes a region in Θ_A on which s_γ is better than content or graph features alone.

Consider $d_\gamma = \mu_\gamma - \max(\mu_1, \mu_2)$ for fixed γ, p_A. Calculate $\frac{\partial d_\gamma}{\partial c_A}$. For $\gamma < 1$, this is non-negative when $\mu_1 > \mu_2$, and non-positive when $\mu_2 > \mu_1$. Thus the maximum value is on curve $\mu_1 = \mu_2$, irrespective of γ. Equivalently the maximum is where test power β_2 equals the β_1 for given p_A.

Consider $D_\gamma = \beta_\gamma - \max\{\beta_1, \beta_2\}$. Below the equal power curve, $D_\gamma = \beta_\gamma - \beta_1$, hence increasing in c_A for fixed p_A. Above the curve, $D_\gamma = \beta_\gamma - \beta_2$. Here taking derivatives along the level curves of β_2 (i.e. fixed $c_A p_A$) shows that D decreases for increasing c_A when $\gamma > 0$. Thus for $\gamma \in (0, 1)$, the maximum for D_γ lies on the equal power curve. ■

Thus the lightest ridge in Fig. 4 is along $\beta_1 = \beta_2$. This illustrates $\gamma = 0.5$, but the theorem implies similar results for $\gamma \in (0, 1)$.

This result does not entirely characterize the performance of the $\lambda_{(p_A, c_A)}$ since some are equivalent to s_γ for $\gamma < 0$. For those, D_γ increases for increasing c_A along fixed $p_A c_A$. Such tests are powerful in the upper left corner of alternative coordinates (p_A, c_A).

The theorem explains our empirical observations of best power improvement on the equal power curve. The gray line of Fig. 5 shows $\mu_1 = \mu_2$, while the black lines are solutions to $\mu_\gamma = \mu_1$ and $\mu_\gamma = \mu_2$. The observed and theoretical lines differ only slightly, suggesting that the normal approximation is good even at $n = 10$.

A classical LR test over Θ_A requires finding the most likely $(\hat{p}_A, \hat{c}_A)$:

$$\begin{aligned}\frac{\partial \lambda}{\partial p_A} &= \frac{\binom{m}{2}(x_1 - \mu_1)}{\sqrt{\binom{n}{2} p_0(1-p_0)}} + \binom{m}{2} \frac{(c_A - c_0)(x_{2|1} - \mu_{2|1})}{\sqrt{t_1 c_0(1-c_0)}} \\ \frac{\partial \lambda}{\partial c_A} &= \binom{m}{2} \frac{p_A (x_{2|1} - \mu_{2|1})}{\sqrt{t_1 c_0(1-c_0)}}.\end{aligned}$$

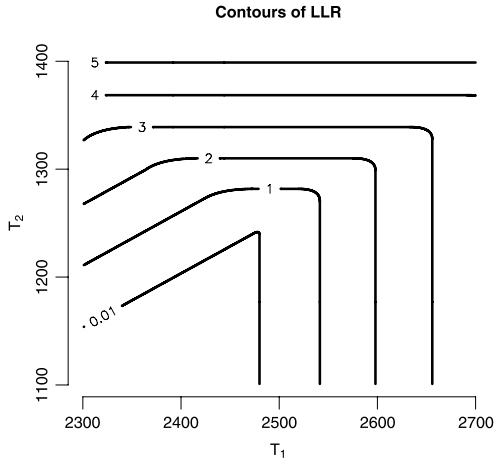

Fig. 7. Contours of log-likelihood ratio, showing shape of rejection regions. Here $n = 100$, $m = 10$, $p_0 = c_0 = 0.5$.

Simultaneously setting both equal to zero results in

$$\hat{p}_A \to \frac{t_1 - \left(\binom{n}{2} - \binom{m}{2}\right) p_0}{\binom{m}{2}}$$

$$\hat{c}_A \to c_0 + \frac{t_2 - c_0 t_1}{a t_1} \to c_0 + \frac{t_2 - c_0 t_1}{\binom{m}{2} p_A}.$$

These can be solved iteratively for m known. For the ratio $m/\sqrt{n}$ unknown, the problem is under-determined since larger $\hat{m}$ can be balanced by smaller $\hat{p}_A$ and $\hat{c}_A$. Thus a range of solutions exists setting $x_1 = \mu_1$ and $x_{2|1} = \mu_{2|1}$, so long as bounds $\hat{c}_A \in [c_0, 1]$ and $\hat{p}_A \in [p_0, 1]$ are respected.

Assuming m known, the parameter restrictions still complicate analysis. In the case where no constraints are invoked, $\hat{\mu}_1 = x_1$ and $\hat{\mu}_{2|1} = x_{2|1}$:

$$\lambda(t_1, t_2) = \frac{x_1^2}{2} + \frac{x_{2|1}^2}{2} = \frac{\left(t_1 - \binom{n}{2} p_0\right)^2}{\binom{n}{2} 2p_0(1-p_0)} + \frac{(t_2 - c_0 t_1)^2}{2t_1 c_0 (1-c_0)}. \tag{8}$$

Eight special cases exist in which the ML estimates lie on the boundary; each can be entered into Eq. (7). Cases $\hat{c}_A = c_0$ lead to tests that reject H_0 for large t_1. Cases $\hat{p}_A = p_0$ lead to tests that reject for large $t_2 - c_0 t_1$. If both hold, the LR is zero and there is no evidence to reject. Estimates $\hat{p}_A = 1$ and/or $\hat{c}_A = 1$ change the form of the LLR but do not greatly simplify it.

Eq. (8) is a test on the location parameter of a multivariate normal distribution, although the covariance on the second term is decidedly non-standard. Large values of $t_1 - \binom{n}{2} p_0$ or $t_2 - c_0 t_1$ are evidence for rejection of the null hypothesis — Fig. 7 shows contours in the (T_1, T_2) plane. There is a literature on generalizations of one-sided normal hypothesis tests to the multivariate case, and Eq. (8) resembles the LR test of Kudo (1963) or the simplified chi-square test of Follmann (1996). The paper Chongcharoen et al. (2002) studies the powers of several related tests, with Kudo's original test typically performing as well or better than proposed alternatives.

While no best test exists over all of Θ_A, the literature suggests that the standard LR test performs well. Fig. 8 displays power contours of tests based upon $T_1, T_2, s_{0.5}$, and λ. In terms of average power (flat distribution on Θ_A), $T_1 < s_{0.5} < \lambda < T_2$, but this depends on the particular choice of null parameters (p_0, c_0). T_1 does best in the lower right corner of Θ_A, but rather poorly compared to the others overall. T_2 does best along the top, with the fusion tests doing well in the intermediate region. Fig. 9 compares the powers of the fusion tests. Fig. 10 compares the powers of individual $\lambda_{(p_A, c_A)}$ to the global LR test. Despite highly variable ML parameter estimates based upon one observed graph, the power difference from the null level $\beta_\lambda - \alpha$ averages more than 90% that of $\max\{\beta_1, \beta_2\} - \alpha$, and more than 85% that of the theoretical ceiling using optimal $\lambda_{(p_A, c_A)}$ at each point.

4. Results on more general graph models

Several generalizations of our basic hypothesis test seem desirable. Here we briefly consider the effects of certain model extensions on the results of the previous section.

4.1. More than 2 topics

Extending one fixed topic of interest to multiple mutually exclusive topic labels with global relative probabilities is straightforward. Bernoulli parameter c becomes probability vector $\vec{c}$. T_2 becomes a vector of topic counts, following

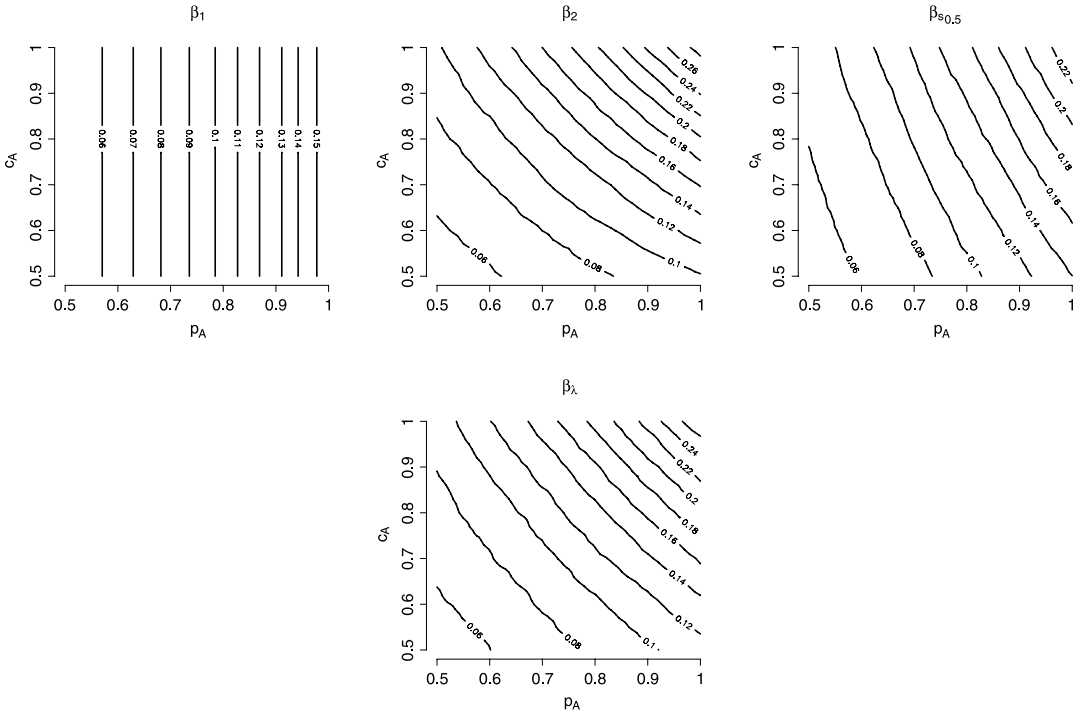

Fig. 8. Empirical power results using T_1, T_2, $s_{0.5}$, and λ on simulation data. Here $(n, m, p_0, c_0, \alpha) = (100, 10, 0.5, 0.5, 0.05)$.

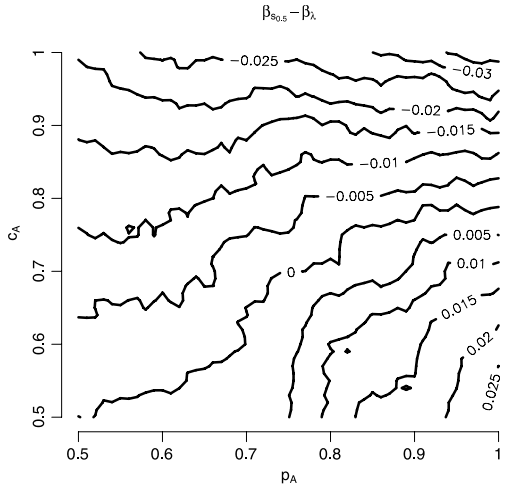

Fig. 9. Power difference $\beta_{0.5} - \beta_\lambda$ on simulation data.

a multinomial rather than binomial distribution. Asymptotically this leads to Gaussian (T_1, T_2) with known mean and covariance, with the earlier analysis essentially unchanged.

4.2. Heterogeneous H_0

A real-world communication graph bears little resemblance to an ER random graph. What follows from relaxing the null assumption of global edge probability p_0? Most generally, we can define local p_{ij} between vertices i and j. This allows a null matrix P_0 of background message probabilities. We denote the set of potential edges among the V^A by E_A. Thus we have null $\mathrm{ER}_c(n, P^0, c_0)$ versus alternative $\mathcal{K}_c(n, P^0, c_0, m, P^A, c_A)$ where $m \times m$ matrix P^A encodes alternate edge probabilities in E_A. The natural interpretation of the null is now lack of change rather than homogeneity; interest is in detecting that communications among some group have changed with respect to previously observed levels.

Matrix P can encode a sparse structure via zero probabilities. In such a case $\mathbb{E}(T_1) = O(n^2)$ or $\mathbb{E}(T_1^A) = O(m^2)$ are not guaranteed. For the earlier observed asymptotic behavior to follow, rather than $m = O(n^{0.5})$ we directly impose the constraint $\mathbb{E}(T_1^A) = O(\mathbb{E}(T_1)^{0.5})$.

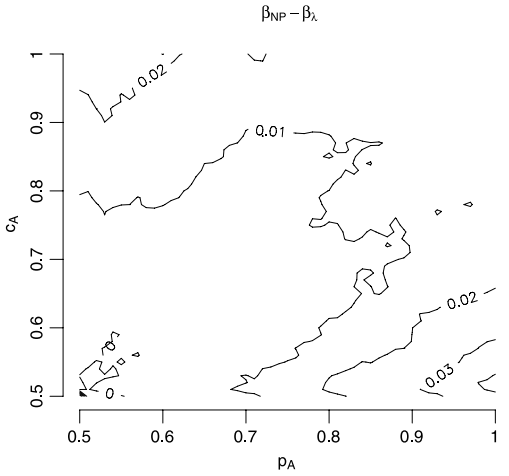

Fig. 10. Power difference of $\lambda_{(p_A,c_A)}$ versus λ test powers on simulation data.

The membership of V^A now impacts mean and variance under H_A. Without restricting the V^A, there are $\binom{n}{m}$ possibilities. It becomes necessary to analytically or numerically average over these possibilities to determine global statistics. For simplicity, assume fixed entries of P^A. Assuming all choices of V^A equally likely,

$$\begin{aligned}\mathbb{E}T_1 &=_{H_0} \sum_{i<j}^{n} P_{ij}^0 \\ &=_{H_A} \left(1 - \frac{\binom{m}{2}}{\binom{n}{2}}\right) \sum_{i<j}^{n} P_{ij}^0 + \sum P_{ij}^A.\end{aligned}$$

Then

$$\begin{aligned}\mathrm{Var}(T_1) &=_{H_0} \sum_{i<j}^{n} P_{ij}^0(1 - P_{ij}^0) \\ &=_{H_A} \left(1 - \frac{\binom{m}{2}}{\binom{n}{2}}\right) \sum_{i<j}^{n} P_{ij}^0(1 - P_{ij}^0) + \sum P_{ij}^A(1 - P_{ij}^A).\end{aligned}$$

Results on conditional $T_2 \mid T_1$ hold as before, but now

$$a(t_1) = \frac{1}{t_1}\left(\sum P_{ij}^A + \frac{\sum P_{ij}^A(1 - P_{ij}^A)\left(t_1 - \sum P_{ij}^A - \left(1 - \frac{\binom{m}{2}}{\binom{n}{2}}\right)\sum_{i<j}^{n} P_{ij}^0\right)}{\sum P_{ij}^A(1 - P_{ij}^A) + \left(1 - \frac{\binom{m}{2}}{\binom{n}{2}}\right)\sum_{i<j}^{n} P_{ij}^0(1 - P_{ij}^0)}\right).$$

While Bernoulli trials corresponding to each edge are no longer identically distributed, asymptotic analysis similar to our previous results holds by the Generalized Central Limit Theorem,

The intuition of V^A representing increased communications about some topic becomes less clear. Should we insist that $\min\{P_{ij}^A\} \geq \max\{P_{ij}^0\}$? Perhaps the entries of P^A should be some function of the entries in P^0 corresponding to the particular choice of V^A. However, averaging functional increases over all possible V^A may not be computationally feasible.

Using local topic probabilities $\overrightarrow{c_{ij}}$ seems problematic in general. Marginal analysis on T_2 can be made as on T_1 using P_{ij} above. However, the conditional distribution of T_2 given T_1 becomes difficult to evaluate, since T_1 could provide information about the membership of anomalous set V^A.

4.3. Multigraphs

Another generalization is to allow $G = (V, A)$ to be a multigraph. For some applications (e.g. email networks Priebe et al., 2005), communications naturally divide into distinct messages. Beyond recognizing that communication took place, one can determine how many, and recognize features of individual messages. Including multiple edges between entities may be a more appropriate model than a simple random graph.

For $n \geq 2$ and $c_0 \in [0, 1]$, and $m \in \{2, \ldots, n-2\}$ and $c_A \in [0, 1]$ with $\lambda_A > \lambda_0 > 0$, $c_A > c_0$, let the null hypothesis H_0 be a random Poisson multigraph $\mathcal{M}_c^{\mathcal{P}}(n, \lambda_0, c_0)$ with Poisson(λ_0) edges between each pair of vertices. H_A, denoted

$\mathcal{K}_c^{\mathcal{P}}(n, \lambda_0, c_0, m, \lambda_A, c_A)$, is defined as previous $\mathcal{K}_c(n, p_0, c_0, m, p_A, c_A)$ using Poisson random variables to generate random multi-edges between vertex pairs, with each edge having a random topic attribute. $\Theta_A = [\lambda_0, \infty) \times [c_0, 1] \setminus \{\lambda_0, c_0\}$.

From the additivity and binomial thinning property of Poisson distributions,

$$\begin{aligned} T_1 &\sim_{H_0} \text{Poisson}\left(\binom{n}{2}\lambda_0\right) \\ &\sim_{H_A} \text{Poisson}\left(\binom{n}{2}\lambda_0 + \binom{m}{2}(\lambda_A - \lambda_0)\right) \\ T_2 &\sim_{H_0} \text{Poisson}\left(\binom{n}{2}c_0\lambda_0\right) \\ &\sim_{H_A} \text{Poisson}\left(\binom{n}{2}c_0\lambda_0 + \binom{m}{2}(c_A\lambda_A - c_0\lambda_0)\right). \end{aligned}$$

Under H_A, conditioning on the sum of independent Poisson variables leads to

$$T_1^A \mid T_1 = t_1 \sim \text{Bin}\left(t_1, \frac{\binom{m}{2}\lambda_A}{\binom{m}{2}(\lambda_A - \lambda_0) + \binom{n}{2}\lambda_0}\right)$$

$$a = \frac{\binom{m}{2}\lambda_A}{\binom{m}{2}(\lambda_A - \lambda_0) + \binom{n}{2}\lambda_0}.$$

Asymptotic results using the Central Limit Theorem can be derived as in Section 3. In particular,

$$\begin{aligned} \lambda(t_1, t_2) = t_1 &\left[\frac{\binom{m}{2}(\lambda_A - \lambda_0)}{\binom{n}{2}\lambda_0} - \frac{a(c_A - c_0)}{1 - c_0} - \frac{a^2(c_A - c_0)^2}{2c_0(1 - c_0)}\right] \\ &+ t_2\frac{a(c_A - c_0)}{c_0(1 - c_0)} - \binom{m}{2}(\lambda_A - \lambda_0)\left[1 + \frac{\binom{m}{2}(\lambda_A - \lambda_0)}{2\binom{n}{2}\lambda_0}\right] \end{aligned} \tag{9}$$

with ML estimates

$$\hat{\lambda}_A = \lambda_0 + \frac{t_1 - \binom{n}{2}\lambda_0}{\binom{m}{2}}$$

$$\hat{c}_A = c_0 + \frac{t_2 - c_0 t_1}{a t_1}$$

subject to constraints $\lambda_A > 0$, $c_A \in [p_0, 1]$.

Corollary 4.1. *Asymptotically as $n \to \infty$, for almost any (p_A, c_A), there exists a test using (T_1, T_2) that is more powerful at that point than a test based on T_1 or T_2 alone. No UMP test exists for $H_0 : \mathcal{M}_c^{\mathcal{P}}(n, \lambda_0, c_0)$ versus $H_A : \mathcal{K}_c^{\mathcal{P}}(n, \lambda_0, c_0, m, \lambda_A, c_A)$ over all Θ_A. There exist curves in Θ_A on which a UMP test using (T_1, T_2) exists.*

Proof. Same argument as for Section 2. ■

5. Conclusions

We have demonstrated that statistical inference on random graphs with attributes associated with the graph features and content of communications can, under simple hypotheses, benefit from fusion of the disparate information types. That benefit is not assured, since incorporating a weak feature can decrease power. Fusion performs best in those regions where individual tests have similar power. For a particular inference problem, we have derived a maximum likelihood ratio test with overall discrimination above the null level on the space of alternatives about 90% of that from the maximum of the best tests based on graph features alone or content alone.

Acknowledgement

The authors would like to thank Charles Wayne for suggesting the importance of communication graphs in human language technology.

References

Airoldi, E.M., Blei, D.M., Feinberg, S.E., Xing, E.P., 2008. Mixed membership stochastic blockmodels. Journal of Machine Learning Research 9, 1823–1856.
Bollobas, B., 2001. Random Graphs. Cambridge University Press.
Chongcharoen, S., Singh, B., Wright, F.T., 2002. Powers of some one-sided multivariate tests with the population covariance matrix known up to a multiplicative constant. Journal of Statistical Planning and Inference 107 (1–2), 103–121.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R., 2004. The Automatic Content Extraction (ACE) program tasks, data, and evaluation. In: Proc. LREC, pp. 837–840.
Follmann, D., 1996. A simple multivariate test for one-sided alternatives. Journal of the American Statistical Association 91, 854–861.
Gilbert, E.N., 1959. Random graphs. Annals of Mathematical Statistics 30 (4), 1141–1144.
Grothendieck, J., Gorin, A., Borges, N., 2009. Social correlates of turn-taking behavior.
Kudo, A., 1963. A multivariate analogue of the one-sided test. Biometrika 50, 403–418.
Leenders, R.Th.A.J., 1995. Structure and Influence: Statistical Models for the Dynamics of Actor Attributes, Network Structure, and their Interdependence. Thesis Publishers.
Priebe, C.E., Conroy, J.M., Marchette, D.J., Park, Y., 2005. Scan statistics on Enron graphs. Computational and Mathematical Organization Theory 11 (3), 229–247.
Sen, S., Spatscheck, O., Wang, D., 2004. Accurate, scalable in-network identification of P2P traffic using application signatures. In: Proc. WWW'04, pp. 512–521.