Graph Inference with Imperfect Edge Classifiers

Michael W. Trosset

Department of Statistics

Indiana University

Joint work with David Brinda (Yale University) and Shantanu Jain (Indiana University), supported by a National Security Science & Engineering Faculty Fellowship awarded to Carey E. Priebe (Johns Hopkins University).

The Smartest Guys in the Room

FERC posted 1.5 million email messages from Enron users. This corpus suffered from document integrity problems and included sensitive/private information. CMU distributed an improved corpus of $517,431$ messages from 150 Enron users (184 email addresses) in 1999–2002 (189 weeks). A subset of 5000 messages sent in 2001 were manually indexed by Murray Browne & Ben Signer and partitioned into the following topics:

CA-analysis (304)	Daily-business (1595)	Downfall-newsfeed (48)	9-11 (29)
CA-bankruptcy (36)	Education (92)	Broadband (26)	9-11-Analysis (30)
CA-utilities (116)	EnronOnline (271)	Federal-gov (85)	Dynegy (7)
CA-crisis-legal (109)	Kitchen-daily (37)	FERC-DOE (219)	Sempra (16)
CA-enron (699)	Kitchen-fortune (11)	College Football (100)	Duke (17)
CA-federal (61)	Energy-newsfeed (332)	Pro Football (6)	El Paso (34)
Newsfeed-CA (190)	General-newsfeed (48)	India-General (38)	Pipelines (17)
CA-legis (181)	Downfall (158)	India-Dabhol (79)	World-energy (25)

Priebe and collaborators have proposed scan statistics for detecting anomalies in time-evolving graphs and hypergraphs. Grothendieck, Priebe, and Gorin (2010) developed some theory for a much simpler task, recently extended by Brinda, Jain, and Trosset (2011).

Experiments on Random Graphs

Consider a random graph with ν vertices. An edge connects vertices i & j with unknown probability π_{ij}. An edge possesses attribute k with unknown conditional probability c_k. We test simple hypotheses about (π, c).

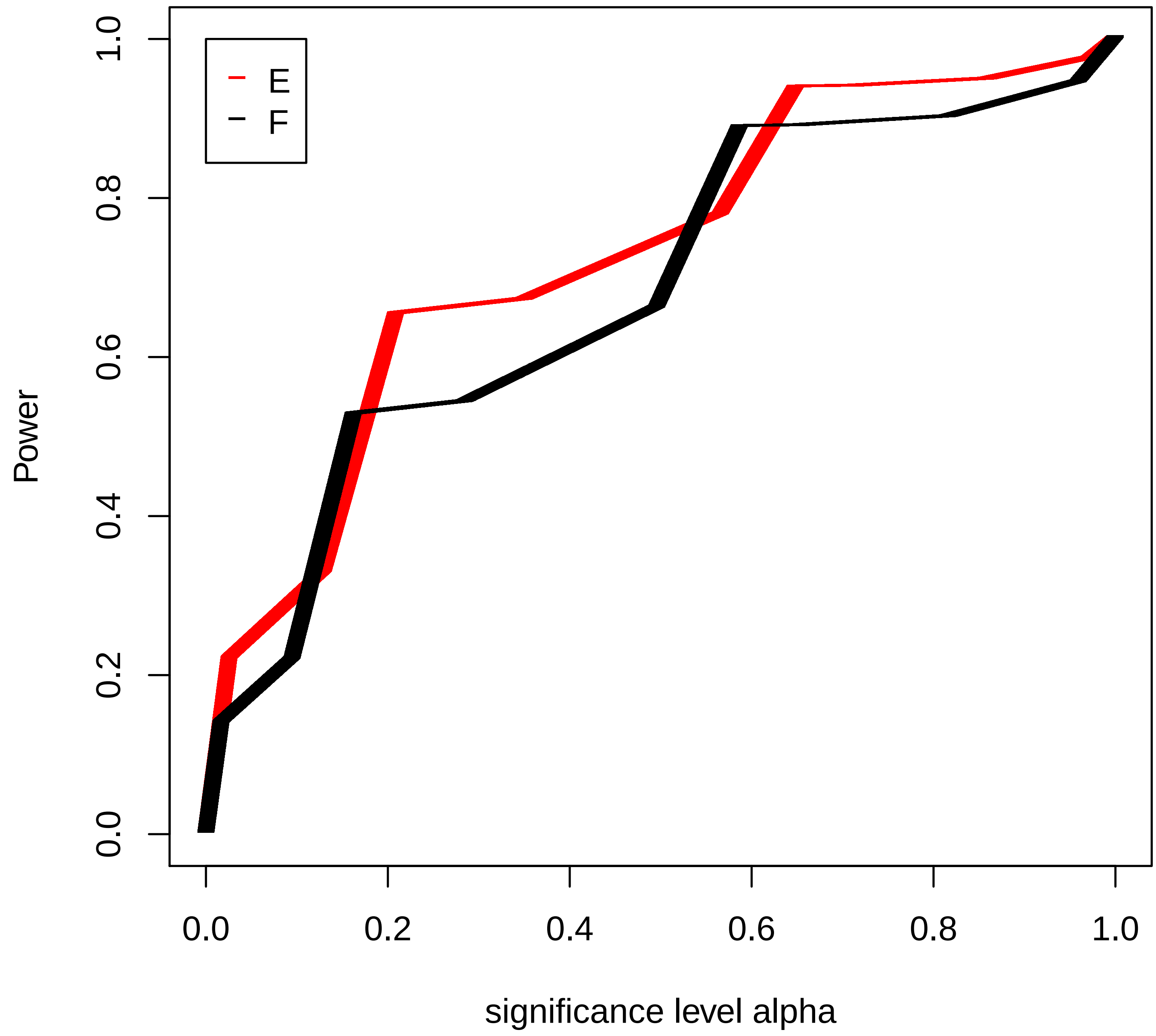
Assuming that the attributes of the edges are known, GPG derived the most powerful test, ϕ_*, for the special case of an Erdös-Renyi graph ($\pi_{ij} = \bar{\pi}$).

We observe edges, but not attributes. Instead, we observe output from a fallible classifier with known confusion matrix $E = [e_{k\ell}]$, where $e_{k\ell}$ is the probability that an edge of type k will be classified as an edge of type ℓ. Note that E is stochastic.

Using classified edge attributes degrades the performance of ϕ_*. How is the performance of ϕ_* affected by the performance of the classifier?

Example (Shantanu Jain): $\nu = 3$,
$H_0 : (\bar{\pi}, c) = (0.60, 0.65)$ vs $H_1 : (\bar{\pi}, c) = (0.90, 0.95)$,

$$E = \begin{bmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}$$


E
F
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
significance level alpha

MP Tests with Fallible Classifiers

For simple hypotheses, the most powerful (MP) test can be determined by application of the Neyman-Pearson Lemma.

Note that the MP test depends on the classifier. In particular, if the classifier is fallible then ϕ_* is not MP.

Given a significance level α, we denote the MP level-α test with a classifier having confusion matrix E by $\phi_E(\cdot;\alpha)$, and the corresponding probability of a Type II error by $\beta_E(\alpha)$.

We investigate how the performance of the classifier affects the power of the test.

Do Better Classifiers Entail Better Tests?

One must be careful about how one compares classifiers. A classifier with

$$F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{is just as good as a classifier with} \quad E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

because a test based on F can simply reverse the attribute assignments and then proceed as though it was based on E.

Suppose that $e_{kk} \geq e_{k\ell}$ and $f_{kk} \geq f_{k\ell}$. For such classifiers, we define the following partial ordering of confusion matrices:

$$E \succ F \quad \text{if and only if} \quad e_{k\ell} \leq f_{k\ell} \; \forall \; k \neq \ell$$

We undertook this investigation with the hope of demonstrating that

$$E \succ F \quad \text{entails} \quad \beta_E(\alpha) \leq \beta_F(\alpha).$$

Comparison of Experiments

There are ν vertices, hence $\mu = \nu(\nu - 1)/2$ possible edges and $N = (K + 1)^{\mu}$ possible outcomes. Our experiment (testing simple hypotheses using E) is completely characterized by a $2 \times N$ stochastic matrix P.

Following Blackwell & Girshick (1954), the experiment P is more informative than the experiment Q ($P \supset Q$) iff there exists a stochastic matrix M such that $PM = Q$. Furthermore, $P \supset Q$ iff for every significance level α the MP level-α test for P is more powerful than than the MP level-α test for Q.

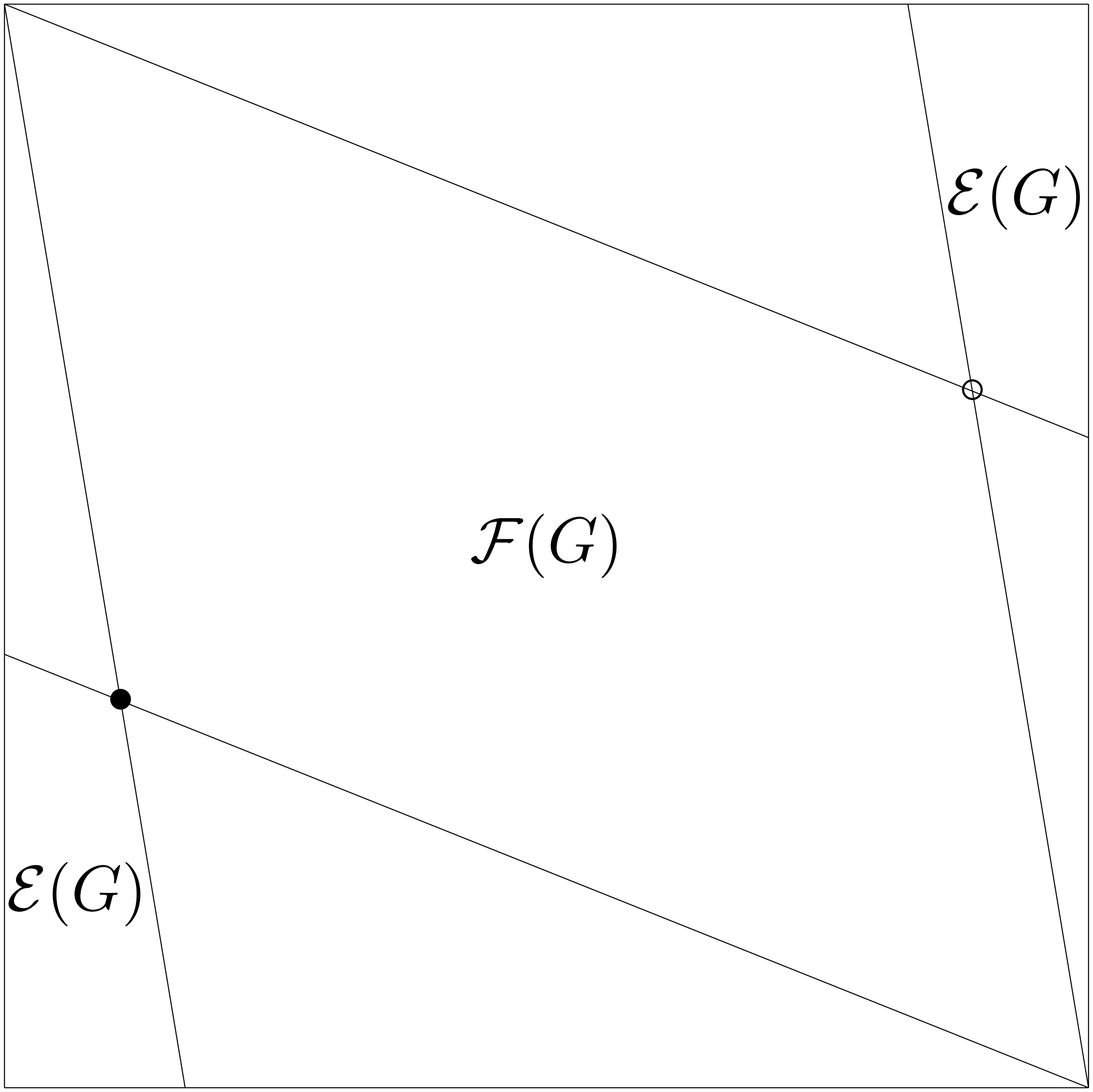
Theorem: If $E \supset F$, then $\beta_E(\alpha) \leq \beta_F(\alpha)$ for every $\alpha \in [0, 1]$.

Proof: Let P and Q denote the experiments associated with E and F. Then $P\left(BU_1BU_2 \cdots BU_{\mu}\right) = Q$, where B is a block diagonal stochastic matrix having blocks of

$$\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & R \end{array}\right]$$

and $U_1, \ldots, U_{\mu}$ are suitable permutation matrices.

Example: $\{\mathbf{F} : \mathbf{G} \supset \mathbf{F}\}$ and $\{\mathbf{E} : \mathbf{E} \supset \mathbf{G}\}$



$K = 2$ edge attributes, error probabilities $g_{12} = 3/28$ and $g_{21} = 10/28$.

Counterexample to Conjecture

Consider the confusion matrices

$$E = \begin{bmatrix} 0.5 & 0.1 & 0.4 \\ 0.1 & 0.5 & 0.4 \\ 0.3 & 0.3 & 0.4 \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} 0.48 & 0.11 & 0.41 \\ 0.10 & 0.50 & 0.40 \\ 0.30 & 0.30 & 0.40 \end{bmatrix}.$$

There is no R for which $ER = F$.

Suppose that $\pi_{ij} = \bar{\pi}$ and we test $H_0 : (\bar{\pi}, c) = (6, 4, 4, 4)/12$ versus $H_1 : (\bar{\pi}, c) = (9, 2, 2, 8)/12$ using either E or F. For $\nu = 2$, the experiments are characterized by the stochastic matrices

$$P = \begin{bmatrix} 40 & 12 & 12 & 24 \\ 20 & 9 & 9 & 24 \end{bmatrix} /80 \quad \text{and} \quad Q = \begin{bmatrix} 1200 & 352 & 364 & 484 \\ 600 & 534 & 543 & 723 \end{bmatrix} /2400.$$

There does not exist a stochastic matrix M for which $PM = Q$. Hence, P is not more informative than Q and therefore there exists $\alpha \in [0, 1]$ for which $\beta_E(\alpha) > \beta_F(\alpha)$.

References and Acknowledgments

J. Grothendieck, C.E. Priebe, A.L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content. *Computational Statistics and Data Analysis*, 54(7):1777–1790, 2010.

W.D. Brinda, S. Jain, M.W. Trosset. *Inference on Random Graphs with Classified Edge Attributes*. Technical Report 11-03, Department of Statistics, Indiana University, October 2011.

D. Blackwell, M.A. Girshick. *Theory of Games and Statistical Decisions*. John Wiley & Sons, 1954.

David Brinda earned an M.S. in Applied Statistics at Indiana University and is currently a Ph.D. student in Statistics at Yale University.

Shantanu Jain is currently a Ph.D. student in Computer Science at Indiana University.

Special thanks to Dave Marchette and the inimitable Carey Priebe—15 years since Interface 1997 in Houston, TX!

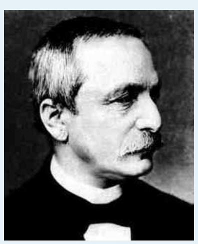
Kronecker Quote

"The wealth of your practical experience

with sane and interesting problems

will give to mathematics

a new direction and a new impetus."



– Leopold Kronecker to Hermann von Helmholtz –

Paraphrased Quote

The wealth of Carey's practical experience

with sane and interesting problems

has given to Michael's mathematics

a new direction and a new impetus.