Coherent vs. Incoherent Graph Classification

Henry Pao, Carey E. Priebe

January 25, 2011

Abstract

1 Introduction

We consider the classification of (labeled) graphs. A random graph $G = (V, E)$, with $V = [n] = \{1, 2, \dots n\}$. These graphs are simple direct graphs with loops. Thus the adjacency matrix has n^2 entries of interest.

Consider $\{(G_i, Y_i)\}_{i=1}^{s} \overset{iid}{\sim} F_{GY}$, with class labels $Y : \Omega \to \{0, 1\}$ and graphs $G : \Omega \to \mathcal{G}_n$, where $\mathcal{G}$ denotes the collection of simple directed graphs with loops. For simplicity, we assume that the prior probability of class membership $\pi = P[Y = 1]$ is known to be 1/2. and the sample sizes $S_y = \sum_{i=1}^{s} I\{Y = y\}$ are fixed. Thus $s_0 = s_1 = s/2$. We consider the independent edge model (IE), so that for $y \in \{0, 1\}$ the class-conditional distribution $F_{G|Y=y}$ is parameterized by an $n \times n$ matrix with entries $p_{y;u,v} \in [0, 1]$.

1.1 IE1-star graph

In order to study the importance of coherence, we designed a special graph whose coherence is easy to take advantage of. Except for m edges this graph has the same distribution as an Erdos-Renyi graph. To chose these m edges first a vertex v^* is uniformly chosen out of the n verticies. Now m edges out of $2n - 1$ edges containing this vertex are chosen to have an edge probability of q.

1.2 Classifier

The Bayes optimal classifier for obseved graph G is

$$g^*(G) = \arg\max_y \prod_{(u,v) \in V \times V} f(a_u, v; p_{y;u,v}),$$

where the Bernoulli probability $f(a; p) = pI\{a = 1\} + (1 - p)I\{a = 0\}$.

The independent edge homogeneous vs inhomogeneous model (IE1), parameterized by n, p, q, and $\mathcal{E} \subset V \times V$, is given by $p_{0;u,v} = p$ for all $(u, v) \in V \times V$ and $p_{1;u,v} = q$ for all $(u, v) \in \mathcal{E}, p_{1;u,v} = p$ for all $(u, v) \in (V \times V) \backslash \mathcal{E}$; $\mathcal{E}$ is the collection of signal edges and $\mathcal{E}^c = (V \times V) \backslash \mathcal{E}$ is the collection of noise edges. (Notice that $F_{G|Y} = 0$ is Erdos-Renyi ER(n, p).) In this model, all signal edges are created equally, and all noise edges are created equally; we will see that this property simplifies our analysis.

In IE1, only $\mathcal{E}$ is relevant and g can be written as

$$g^*(G) = \arg\max_y \prod_{(u,v) \in \mathcal{E}} f(a_{u,v}; p_{y;u,v}).$$

If we estimate $p_{y;u,v}$ from the training data, we may consider classifiers

$$g_{NB}(G) = \arg\max_y \prod_{(u,v) \in V \times V} f(a_{u,v}; p_{y;u,v})$$

and

$$g_{\mathcal{E}}(G) = \arg\max_y \prod_{(u,v)\in\mathcal{E}} f(a_{u,v}; p_{y;u,v}).$$

The latter is the best we can hope for it considers the signal edges and only the signal edges; the former can be swamped by noise from non-signal edges.

Our interest is canonical subspace identification for this graph classification application; that is, estimate the collection of signal edges $\mathcal{E}$ via $\hat{\mathcal{E}}$ and consider the classifier

$$g_{\hat{\mathcal{E}}}(G) = \arg\max_y \prod_{(u,v)\in\hat{\mathcal{E}}} f(a_{u,v}; p_{y;u,v}).$$

We consider two different methods to estimate $\hat{\mathcal{E}}$ for IE1-star graphs.

if $q > p$, let $\delta_{u,v} = p_{1;u,v} - p_{0;u,v}$. thus $\hat{\delta}_{u,v} = \hat{p}_{1;u,v} - \hat{p}_{0;u,v}$

1.2.1 incoherent method: agnostic

The incoherent method does not utilize the stucture of the graph. Let the number of signal edges we will attempt to extract be $k = |\hat{\mathcal{E}}|$. Then our incoherent model is the k largest $\hat{\delta}_{u,v}$ edges. For simplicity, the rest of this paper we will instead use $\hat{q}_{u,v}$ because this simplifies our theoretical calcuations.

1.2.2 coherrent method: max degree

This method takes advanage of the fact the IE1-star graphs has a vertex v^* which all edges with probability q are adjacent to. For convience let $v \in (u_1, u_2) \in V \times V$ mean $(u_1, u_2) \in V \times V$, and $u_1 = v$ or $u_2 = v$ (or $u_1 = u_2 = v$). First the coherent method estimates this vertex

$$\hat{v}^* = \arg\max_v \sum_{v\in(u_1,u_2)\in V\times V} \hat{\delta}_{u_1,u_2}$$

$\hat{\mathcal{E}}$ is the k largest $\hat{\delta}_{u,v}$ edges adjecent to v^*. Ideally we would like to use δ_{u_1,u_2}; however, because it highly complicates theoretical calculations, we will instead use $\hat{q}_{u_1,u_2}$.

2 Theoretical results

2.1 Monotonisity of error given T

In IE1, using k canonical dimensions recovered from the training data ($|\hat{\mathcal{E}}| = k$), the probability of misclassification is monotonically decreasing as a function of $T = |\mathcal{E} \cap \hat{\mathcal{E}}|$ that is

$$t_1 > t_2 \Rightarrow E[L(g_{\hat{\mathcal{E}}})|T = t_1] < E[L(g_{\hat{\mathcal{E}}})|T = t_2].$$

2.1.1 $k = 1$ case

First consider the case where only one signal edge is attempted to be recovered ($k = 1$). Let g_0 represent the classifer if the recovered edge is not a signal edge ($t = 0$) and g_1 represent the classifier if the recovered edge is a signal edge ($t = 1$). If the above montonisity result is true we expect

$$E[L(g_1)] < E[L(g_0)].$$

Since we only have one edge, for simplicity let $\hat{p}_0$ and $\hat{p}_1$ denote the estimates of p_0 and p_1 for our recovered edge respectively. The following decomposes $E[L(g_0)]$ using the law of total probability conditioning on a, Y.

$$\begin{aligned}
E[L(g_0)] &= P[g_0 \neq Y] = P[\arg\max_y f(a;\hat{p}_y) \neq Y] && (1)\\
&= \sum_{j\in\{0,1\}} P[Y=j]P[\arg\max_y f(a;\hat{p}_y) \neq Y|Y=j] && (2)\\
&= \sum_{i,j\in\{0,1\}} P[Y=j]P[a=i|Y=j]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=i,Y=j] && (3)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=0,Y=0] && (4)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=1,Y=0] && (5)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=0,Y=1] && (6)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(a;\hat{p}_y) \neq Y|a=1,Y=1] && (7)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(0;\hat{p}_y) \neq 0] && (8)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(1;\hat{p}_y) \neq 0] && (9)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(0;\hat{p}_y) \neq 1] && (10)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(1;\hat{p}_y) \neq 1] && (11)\\
&= \frac{1}{2}(1-p)P[\arg\max_y(1-\hat{p}_y) \neq 0] && (12)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 0] && (13)\\
&\quad +\frac{1}{2}(1-p)P[\arg\max_y(1-\hat{p}_y) \neq 1] && (14)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 1] && (15)
\end{aligned}$$

Note $\hat{p}_0, \hat{p}_1$ are independent of a, Y. Conditioning on the relationship between $\hat{p}_0$ and $\hat{p}_1$,

$$\begin{aligned}
&= \frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y)) \neq 0|\hat{p}_0 < \hat{p}_1] && (16)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 = \hat{p}_1] && (17)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 > \hat{p}_1]] && (18)\\
&\quad +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 < \hat{p}_1] && (19)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 = \hat{p}_1] && (20)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 > \hat{p}_1]] && (21)\\
&\quad +\frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 < \hat{p}_1] && (22)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 = \hat{p}_1] && (23)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 > \hat{p}_1]] && (24)\\
&\quad +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 < \hat{p}_1] && (25)\\
&\qquad +P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 = \hat{p}_1] && (26)\\
&\qquad +P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 > \hat{p}_1]] && (27)
\end{aligned}$$

In the event that $\hat{p}_0 = \hat{p}_1$ the classifier's decicion is randomized thus $P[g_y \neq Y|\hat{p}_0 = \hat{p}_1] = \frac{1}{2}$ [true??] for $y = \{0, 1\}$. Notice with the conditional probabilities relating to g_y are either 0, 0.5, or 1.

$$\begin{aligned}
&= \frac{1}{2}(1-p)[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (28)\\
&\quad +\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (29)\\
&\quad +\frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] && (30)\\
&\quad +\frac{1}{2}p[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] && (31)\\
&= \frac{1}{2}P[\hat{p}_0 = \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}p + \frac{1}{2}(1-p) + \frac{1}{2}p] && (32)\\
&\quad +\frac{1}{2}(1-p)P[\hat{p}_0 > \hat{p}_1] + \frac{1}{2}pP[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}(1-p)P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}pP[\hat{p}_0 > \hat{p}_1] && (33)\\
&= \frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 > \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 < \hat{p}_1] && (34)\\
&= \frac{1}{2} && (35)
\end{aligned}$$

Now consider the classifier g_1. Repeating the same conditioning as on g_0 we get

$$\begin{aligned}
E[L(g_1)] &= P[g_1 \neq Y] && (36)\\
&= P[Y=0]P[a=0|Y=0]P[\arg\max_y f(0;\hat{p}_y) \neq 0] && (37)\\
&\quad +P[Y=0]P[a=1|Y=0]P[\arg\max_y f(1;\hat{p}_y) \neq 0] && (38)\\
&\quad +P[Y=1]P[a=0|Y=1]P[\arg\max_y f(0;\hat{p}_y) \neq 1] && (39)\\
&\quad +P[Y=1]P[a=1|Y=1]P[\arg\max_y f(1;\hat{p}_y) \neq 1] && (40)\\
&= \frac{1}{2}(1-p)P[\arg\max_y (1-\hat{p}_y) \neq 0] && (41)\\
&\quad +\frac{1}{2}pP[\arg\max_y \hat{p}_y \neq 0] && (42)\\
&\quad +\frac{1}{2}(1-q)P[\arg\max_y (1-\hat{p}_y) \neq 1] && (43)\\
&\quad +\frac{1}{2}qP[\arg\max_y \hat{p}_y \neq 1] && (44)
\end{aligned}$$

Now again conditioning on the relationship between $\hat{p}_0$ and $\hat{p}_1$ and noting after conditioning probabilities relating to g_y are either 0, 0.5, or 1.

$$= \frac{1}{2}(1-p)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y)) \neq 0|\hat{p}_0 < \hat{p}_1] \tag{45}$$

$$+P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 = \hat{p}_1] \tag{46}$$

$$+P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 0|\hat{p}_0 > \hat{p}_1]] \tag{47}$$

$$+\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 < \hat{p}_1] \tag{48}$$

$$+P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 = \hat{p}_1] \tag{49}$$

$$+P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 0|\hat{p}_0 > \hat{p}_1]] \tag{50}$$

$$+\frac{1}{2}(1-q)[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 < \hat{p}_1] \tag{51}$$

$$+P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 = \hat{p}_1] \tag{52}$$

$$+P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y(1-\hat{p}_y) \neq 1|\hat{p}_0 > \hat{p}_1]] \tag{53}$$

$$+\frac{1}{2}q[P[\hat{p}_0 < \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 < \hat{p}_1] \tag{54}$$

$$+P[\hat{p}_0 = \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 = \hat{p}_1] \tag{55}$$

$$+P[\hat{p}_0 > \hat{p}_1]P[\arg\max_y \hat{p}_y \neq 1|\hat{p}_0 > \hat{p}_1]] \tag{56}$$

$$= \frac{1}{2}(1-p)[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] \tag{57}$$

$$+\frac{1}{2}p[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] \tag{58}$$

$$+\frac{1}{2}(1-q)[P[\hat{p}_0 < \hat{p}_1] + \frac{1}{2}P[\hat{p}_0 = \hat{p}_1]] \tag{59}$$

$$+\frac{1}{2}q[\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] + P[\hat{p}_0 > \hat{p}_1]] \tag{60}$$

Factoring in terms of $P[\hat{p}_0 < \hat{p}_1]$, $P[\hat{p}_0 = \hat{p}_1]$, and $P[\hat{p}_0 > \hat{p}_1]$.

$$= P[\hat{p}_0 < \hat{p}_1][\frac{1}{2}p + \frac{1}{2}(1-q)] \tag{61}$$

$$+\frac{1}{2}P[\hat{p}_0 = \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}p + \frac{1}{2}(1-q) + \frac{1}{2}q] \tag{62}$$

$$+P[\hat{p}_0 > \hat{p}_1][\frac{1}{2}(1-p) + \frac{1}{2}q] \tag{63}$$

$$= \frac{1}{2}P[\hat{p}_0 < \hat{p}_1][1-(q-p)] \tag{64}$$

$$+\frac{1}{2}P[\hat{p}_0 > \hat{p}_1][1+(q-p)] \tag{65}$$

$$+\frac{1}{2}P[\hat{p}_0 = \hat{p}_1] \tag{66}$$

$$= \frac{1}{2} - (q-p)[P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1]] \tag{67}$$

Remember that $q > p$, so if $P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] > 0$ then $E[L(g_0)] > E[L(g_1)]$. Because our recovered edge is a signal edge, $\hat{p}_0$ ~Binomial(p,s_0) and $\hat{p}_1$ ~Binomial(p,s_1). And since $\hat{p}_0$ and $\hat{p}_1$ are independent,

$$P[\hat{p}_0 > \hat{p}_1] = \sum_{x>y:x,y\in[s/2]} P[\hat{p}_0 = x]P[\hat{p}_1 = y] \tag{68}$$

$$= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x} p^x(1-p)^{s_0-x} \binom{s_1}{y} q^y(1-q)^{s_1-y} \tag{69}$$

$$= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y} \tag{70}$$

Similarly,

$$P[\hat{p}_0 < \hat{p}_1] = \sum_{x>y:x,y\in[s/2]} \binom{s_1}{x}\binom{s_0}{y} q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} \tag{71}$$

Thus

$$P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] = \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y} \tag{72}$$

$$- \binom{s_1}{x}\binom{s_0}{y} q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} \tag{73}$$

Note that $s_0 = s_1$, allowing for us to factor

$$P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] = \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_1}{y} [q^x(1-q)^{s_1-x} p^y(1-p)^{s_0-y} - p^x(1-p)^{s_0-x} q^y(1-q)^{s_1-y}] \tag{74}$$

$$= \sum_{x>y:x,y\in[s/2]} \binom{s_0}{x}\binom{s_0}{y} p^y q^y (1-p)^{s_0-x}(1-q)^{s_0-x}[q^{x-y}(1-p)^{x-y} - p^{x-y}(1-q)^{x-y}] \tag{75}$$

Since $q > p$ and $x - y > 0$, $q^{x-y}(1-p)^{x-y} > p^{x-y}(1-q)^{x-y}$. Therefore $P[\hat{p}_0 < \hat{p}_1] - P[\hat{p}_0 > \hat{p}_1] > 0$ and $E[L(g_1)] < E[L(g_0)]$.

2.1.2 $k > 1$ case

2.2 Assumtotic/Approximate distribution of T

Note that $n\hat{p}_{0;i,j}$ and $n\hat{p}_{1;i,j}$ are binomals, thus they can be approximated with the following normal approximations for n sufficiently large.

$$\hat{p}_{0;i,j} \approx N_{0;i,j} \sim \text{ Normal}(p_{0;i,j}, p_{0;i,j}(1-p_{0;i,j})s/2)$$

$$\hat{p}_{1;i,j} \approx N_{1;i,j} \sim \text{ Normal}(p_{1;i,j}, p_{1;i,j}(1-p_{1;i,j})s/2)$$

2.2.1 Agnostic method

The cumalative distribution function of the r^{th} ordered statistic of n iid random variables is [1]

$$F_{X_{(r:n)}}(x) = P[X_{(r:n)} \leq x] = \sum_{i=r}^{n} \binom{n}{i} F_X(x)^i[1 - F_X(x)]^{n-i}$$

The probability density function of the r^{th} ordered statistic of n iid random variables is [1]

$$f_{X_{(r:n)}}(x) = \binom{n}{r-1, n-r, 1} F_X^{r-1}(x)[1 - F_X(x)]^{n-r} f_X(x)$$

The joint probability density function of 2 ordered statistics $r < s$, $x \leq y$ [1]

$$f_{(r)(s):n}(x,y) = \binom{n}{r-1,1,s-r-1,1,n-s} F^{r-1}(x) f(x) [F(y) - F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s}$$

Let us define another random variable,

$$N_p \sim \text{Normal}(p, p(1-p)2/s)$$

Consider $T = 0$, this occurs when the k^{th} largest nonsignal edge is larger than the largest signal edge.

$$\begin{aligned} P[T=0] &\approx P[N_{p_{(n^2-m-k+1:n^2-m)}} > N_{q_{(m:m)}}] && (76) \\ &= \int f_{N_{p_{(n^2-m-k+1:n^2-m)}}}(x) F_{N_{q_{(m:m)}}}(x)\, dx && (77) \\ &= \int \binom{n^2-m}{n^2-m-k, k-1, 1} F_{N_p}^{n^2-m-k}(x)[1-F_{N_p}(x)]^{k-1} f_{N_p}(x) F_{N_q}(x)^m \, dx && (78) \end{aligned}$$

Now consider when $k \leq m$ $T = k$, this occurs exactly when the k^{th} largest signal edge is greater than the largest nonsignal edge. Since the ordered signal and nonsignal edges are independent,

$$\begin{aligned} P[T=k] &\approx P[N_{q_{(m-k+1:m)}} > N_{p_{(n^2-m:n^2-m)}}] \\ &= \int f_{N_{q_{(m-k+1:m)}}}(x) F_{N_{p_{(n^2-m:n^2-m)}}}(x)\, dx \\ &= \int \binom{m}{m-k, k-1, 1} F_{N_q}^{m-k}(x)[1-F_{N_q}(x)]^{k-1} f_{N_q}(x) F_{N_p}(x)^{n^2-m} \, dx \end{aligned}$$

To be general there can be cases where $k \geq m$ [thus the support of T is $\{0, 1, 2, \ldots, \min(m,k)\}$]. In this case consider when $T = m$, this occurs when the smallest signal edge is greater than $k - m^{th}$ nonsignal edge.

$$\begin{aligned} P[T=m] &\approx P[N_{q_{(1:m)}} > N_{p_{(n^2-m-(k-m):n^2-m)}}] = P[N_{q_{(1:m)}} > N_{p_{(n^2-k:n^2-m)}}] \\ &= \int f_{N_{q_{(1:m)}}}(x) F_{N_{p_{(n^2-k:n^2-m)}}}(x)\, dx \\ &= \int \binom{m}{0, m-1, 1} F_{N_q}^{0}(x)[1-F_{N_q}(x)]^{m-1} f_{N_q}(x) \sum_{i=n^2-k}^{n^2-m} \binom{n^2-m}{i} F_{N_p}(x)^i [1-F_X(x)]^{n^2-m-i} \, dx \\ &= \int m[1-F_{N_q}(x)]^{m-1} f_{N_q}(x) \sum_{i=n^2-k}^{n^2-m} \binom{n^2-m}{i} F_{N_p}(x)^i [1-F_X(x)]^{n^2-m-i} \, dx \end{aligned}$$

Notice that when $k = m$ then the expression for $T = k$ and $T = m$ are the same.

For $t \in [1, 2, \ldots, \min(k,m) - 1]$, the event that $T = t$ is when in the largest k of the δ's t are from a Binomial(q, s_1) and the remaining $k-t$ drawn from Binomial(p, s_0). This means that there are four important ordered statistics, $N_{p_{(n^2-m-k+t:n^2-m)}}$, $N_{p_{(n^2-m-k+t+1:n^2-m)}}$, $N_{q_{(m-t:m)}}$, $N_{q_{(m-t+1:m)}}$. These statistics are at the border of being included in and excluded from $\hat{\mathcal{E}}$. This leads to the expression

$$\begin{aligned}
P[T=t] &\approx P[N_{p_{(n^2-m-k+t:n^2-m)}} < N_{q_{(m-t+1:m)}}, N_{q_{(m-t:m)}} < N_{p_{(n^2-m-k+t+1:n^2-m)}}] \\
&= \iint\limits_{x<y} f_{N_{q_{(m-t)(m-t+1):m}}}(x,y) P[N_{p_{(n^2-m-k+t:n^2-m)}} < y, x < N_{p_{(n^2-m-k+t+1:n^2-m)}}] \; dx \; dy \\
&= \iiiint\limits_{x<y,\ w<y,\ x<z,\ w<z} f_{N_{q_{(m-t)(m-t+1):m}}}(x,y) f_{N_{p_{(n^2-m-k+t)(n^2-m-k+t+1):n^2-m}}}(w,z) \; dw \; dx \; dy \; dz \\
&= \iiiint\limits_{x<y,\ w<y,\ x<z,\ w<z} \binom{m}{m-t-1,1,0,1,t-1} F_{N_q}^{m-t-1}(x) f_{N_q}(x)[F_{N_q}(y)-F_{N_q}(x)]^0 f_{N_q}(y)[1-F_{N_q}(y)]^{t-1} \\
&\quad \binom{n^2-m}{n^2-m-k+t-1,1,0,1,k-t-1} F_{N_p}^{n^2-m-k+t-1}(w) f_{N_p}(w)[F_{N_p}(z)-F_{N_p}(w)]^0 f_{N_p}(z)[1-F_{N_p}(z)]^{k-t-1} \; d \\
&= \iiiint\limits_{x<y,\ w<y,\ x<z,\ w<z} \binom{m}{m-t-1,t-1,1,1} F_{N_q}^{m-t-1}(x) f_{N_q}(x) f_{N_q}(y)[1-F_{N_q}(y)]^{t-1} \\
&\quad \binom{n^2-m}{n^2-m-k+t-1,k-t-1,1,1} F_{N_p}^{n^2-m-k+t-1}(w) f_{N_p}(w) f_{N_p}(z)[1-F_{N_p}(z)]^{k-t-1} \; dw \; dx \; dy \; dz
\end{aligned}$$

Let this distribution be denoted as follows

$$P[T=t] = f_{T_{n^2,m}}(t),$$

where n^2 is the total number of edges, and m is the number of signal edges.

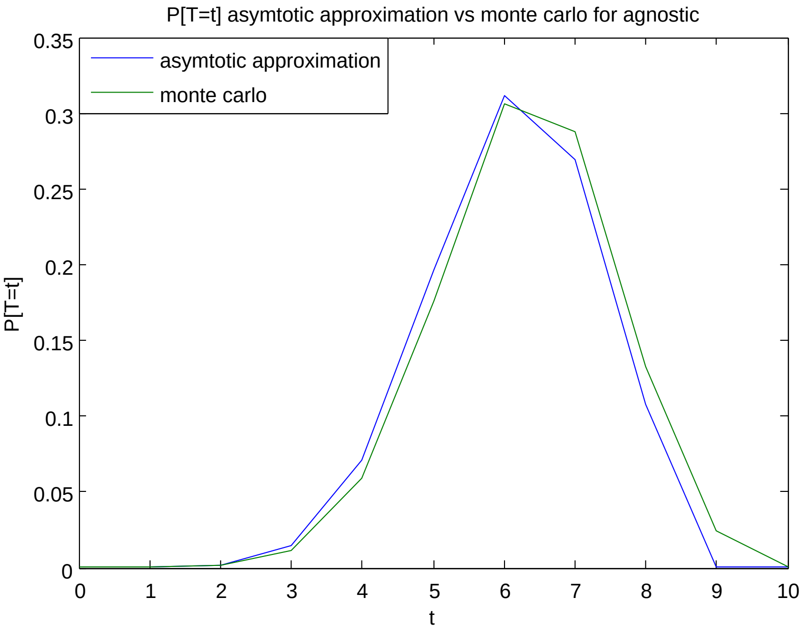

Figure 1: A comparison of a monte carlo simulation and asymtotic distribution of the agnostic method with $n = 10$, $m = 10$, $k = 10$, $p = 0.45$, $q = 0.55$, $s_0 = s_1 = 100$.

Figure 1 shows an example of the accuracy of the asymtotic distribution for a specific set of parameters.

2.2.2 Maximum degree method

Let v^* be the "right" vertex. Let $I = 1$ if the edge (v^*, v^*) is a signal edge [and $I = 0$ if it is not], and R be the number of verticies $u \in V$ such that edges (v^*, u) and (u, v^*) are both signal edges. We will refer to this type of edge as a doubled signal edge. This means that there are exactly R vertices with 2 signal edges, $m - 2R - I$ with 1 signal edge, and $n - 1 - (m - 2R - I) - R = n - m + R - 1 + I$ verticies with 0 signal edges [and v^* with m signal edges]. Then we can rewrite the probability of picking the right vertex as

$$P[\hat{v}^* = v^*] = \sum_{i,r} P[I = i, R = r] P[\hat{v}^* = v^* | I = i, R = r] \tag{86}$$

$P[I = i, R = r]$ can be found with combinitorics. First choose the r doubled signal edges, next choose the vertices for remaining signal edges, and remembering that either incoming or outgoing edge can be signal. We get

$$P[I = 1, R = r] = \frac{\binom{n-1}{r}\binom{n-1-r}{m-2r-1} 2^{m-2r-1}}{\binom{2n-1}{m}} \tag{87}$$

Notice that the support of R depends on m. The support of R is $\max(0, m-n)$ to $\lfloor (m-1)/2 \rfloor$.

Similarly for $I = 0$, except the support of R is now $\max(0, m - n + 1)$ to $\lfloor m/2 \rfloor$ if $m < n + 1$.

$$P[I = 0, R = r] = \frac{\binom{n-1}{r}\binom{n-1-r}{m-2r} 2^{m-2r}}{\binom{2n-1}{m}} \tag{88}$$

Or more consicely for $I \in \{0, 1\}$ and R from $\max(0, m - i - n + 1)$ to $\lfloor (m - i)/2 \rfloor$

$$P[I = i, R = r] = \frac{\binom{n-1}{r}\binom{n-1-r}{m-2r-i} 2^{m-2r-i}}{\binom{2n-1}{m}} \tag{89}$$

Let d_v be the "degree" of vertex v (includes both incoming and outgoing edges)

$$d_v = \sum_{v \in u_1, u_2} \hat{q}_{u_1,u_2}. \tag{90}$$

Given R and I, the event $v \neq v^*$ can be divided into 3 dijoint sets: d_v with 0, 1, or 2 signal edges. Let us denote a vertex with i signal edges as $d(i)$. [Thus $d(2)_{(r:r)}$ is maximum degree of r vertices with 2 signal egdes.]

$$\begin{aligned} P[\hat{v}^* = v^* | I = i, R = r] &= P[d_{v^*} > \max_{v \neq v^*} d_v | I = i, R = r] \\ &= P[d_{v^*} > d(2)_{(r:r)}, d_{v^*} > d(1)_{(m-2r-i:m-2r-i)}, d_{v^*} > d(0)_{(n-m+r-1+i:n-m+r-1+i)}] \end{aligned}$$

Notice that d_{v^*} ~Binomial(q, ms_1)+Binomial$(p, s_1(2n-1-m))$, and d_j ~Binomial$(q, s_1 j)$+Binomial$(p, s_1(2n-$

$1-j)$). Let $I=i$, and using the normal approximation of binomials

$$\begin{aligned}
d(0)_{(n-m+r-1+i:n-m+r-1+i)} &\approx N_{p,(2n-1)_{(n-m+r-1+i:n-m+r-1+i)}}\\
d(1)_{(m-2r-i:m-2r-i)} &\approx \left[N_{p,(2n-2)}+N_{q,1}\right]_{(m-2r-i:m-2r-i)}\\
&= \left[\text{Normal}\left(p(2n-2)+q, \frac{p(1-p)(2n-2)+q(1-q)}{s}\right)\right]_{(m-2r-i:m-2r-i)}\\
d(2)_{(r:r)} &\approx \left[N_{p,(2n-3)}+N_{q,2}\right]_{(r:r)}\\
&= \left[\text{Normal}\left(p(2n-3)+2q, \frac{p(1-p)(2n-3)+2q(1-q)}{s}\right)\right]_{(r:r)}\\
d_{v^*} &\approx N_{q,m}+N_{p,(2n-1-m)}\\
&= \text{Normal}\left(p(2n-1-m)+mq, \frac{p(1-p)(2n-1-m)+mq(1-q)}{s}\right).
\end{aligned}$$

For the convience of notation let the normal approximations be written as

$$\begin{aligned}
d(j) &\approx N_{d(j)}\\
d_{v^*} &\approx N_{d_{v^*}}.
\end{aligned}$$

When $n \to \infty$ the dependence of d_v diminishes []. Thus assumtotically

$$\begin{aligned}
P[\hat{v}^*=v^*|I=i,R=r] &= P[d_{v^*}>d(2)_{(r:r)}, d_{v^*}>d(1)_{(m-2r-i:m-2r-i)}, d_{v^*}>d(0)_{(n-m+r-1+i:n-m+r-1+i)}]\\
&\approx P[d_{v^*}>d(2)_{(r:r)}]P[d_{v^*}>d(1)_{(m-2r-i:m-2r-i)}]P[d_{v^*}>d(0)_{(n-m+r-1+i:n-m+r-1+i)}]
\end{aligned}$$

Note that when $r=0$, $d(2)_{(r:r)}$ does not make sence. $r=0$ means that all verticies but v^* have no more than 2 signal edges. The expression should not have the term $P[d_{v^*}>d(2)_{(r:r)}]$. To rectify this let $F_{d(j)_{(0:0)}}(x)=1$ for all j,x. [This occurs again when $m-2r-i=0$]

Continuing using these normal approximations

$$\begin{aligned}
P[\hat{v}^*=v^*|I=i,R=r] &\approx P[N_{d_{v^*}}>N_{d(2)_{(r:r)}}]P[N_{d_{v^*}}>N_{d(1)_{(m-2r-i:m-2r-i)}}]P[N_{d_{v^*}}>N_{d(0)_{(n-m+r-1+i:n-m+r-1+i)}}]\\
&= \int f_{d_{v^*}}(z)P[z>N_{d(2)_{(r:r)}}]P[z>N_{d(1)_{(m-2r-i:m-2r-i)}}]P[z>N_{d(0)_{(n-m+r-1+i:n-m+r-1+i)}}]\ dz\\
&= \int f_{d_{v^*}}(z)F^{r}_{N_{d(2)}}(z)F^{m-2r-i}_{N_{d(1)}}F^{n-m+r-1+i}_{N_{d(0)}}(z)\ dz
\end{aligned}$$

Now the probability of $\hat{v}^*=v^*$ can be explicitly written

$$\begin{aligned}
P[\hat{v}^*=v^*] &= \sum_{i,r} P[I=i,R=r]P[\hat{v}^*=v^*|I=i,R=r]\\
&\approx \sum_{i,r} \frac{\binom{n-1}{r}\binom{n-1-r}{m-2r-i}2^{m-2r-i}}{\binom{2n-1}{m}} \int f_{d_{v^*}}(z)F^{r}_{N_{d(2)}}(z)F^{m-2r-i}_{N_{d(1)}}F^{n-m+r-1+i}_{N_{d(0)}}(z)\ dz
\end{aligned}$$

Notice given $\hat{v}^*=v^*$, the distribution of T is the same as the agnostic method except instead of n^2-m edges with probability p there are only $2n-1-m$ edges, equation (90). The distribution of $T=t$ for $t>2$ can be approximated as follows

$$\begin{aligned}
P[T=t] &= P[T=t|\hat{v}^*=v^*]P[\hat{v}^*=v^*]+P[T=t|\hat{v}^*\neq v^*]P[\hat{v}^*\neq v^*]\\
&= P[T=t|\hat{v}^*=v^*]P[\hat{v}^*=v^*]\\
&\approx f_{T_{2n-1,m}}(t)\sum_{i,r} \frac{\binom{n-1}{r}\binom{n-1-r}{m-2r-i}2^{m-2r-i}}{\binom{2n-1}{m}} \int f_{d_{v^*}}(z)F^{r}_{N_{d(2)}}(z)F^{m-2r-i}_{N_{d(1)}}F^{n-m+r-1+i}_{N_{d(0)}}(z)\ dz
\end{aligned}$$

For $t \leq 2$ if the wrong vertex is chosen, then it is still possible to correctly pick 0, 1, or 2 signal edges. Using the same method used to calulate $P[\hat{v}^* = v^*|I = i, R = r]$, we can approximate the probability of choosing a vertex with J signal edges

$$\begin{aligned}
P[\hat{v}^* = d(0)|I = i, R = r] &\approx& P[d(0)_{(n-m+r-1+i:n-m+r-1+i)} > d(2)_{(r:r)}] \\
&& P[d(0)_{(n-m+r-1+i:n-m+r-1+i)} > d(1)_{(m-2r-i:m-2r-i)}] \\
&& P[d(0)_{(n-m+r-1+i:n-m+r-1+i)} > d_{v^*}] \\
P[\hat{v}^* = d(1)|I = i, R = r] &\approx& P[d(1)_{(m-2r-i:m-2r-i)} > d(2)_{(r:r)}] \\
&& P[d(1)_{(m-2r-i:m-2r-i)} > d(0)_{(n-m+r-1+i:n-m+r-1+i)}] \\
&& P[d(1)_{(m-2r-i:m-2r-i)} > d_{v^*}] \\
P[\hat{v}^* = d(2)|I = i, R = r] &\approx& P[d(2)_{(r:r)} > d(0)_{(n-m+r-1+i:n-m+r-1+i)}] \\
&& P[d(2)_{(r:r)} > d(1)_{(m-2r-i:m-2r-i)}] \\
&& P[d(2)_{(r:r)} > d_{v^*}]
\end{aligned}$$

For $t \leq 2$ the distribution of T can be written as

$$\begin{aligned}
P[T = t] &=& P[T = t|\hat{v}^* = v^*]P[\hat{v}^* = v^*] + \sum_{j=0}^{2} P[T = t|\hat{v}^* = d(j)]P[\hat{v}^* = d(j)] \\
&=& P[T = t|\hat{v}^* = v^*]P[\hat{v}^* = v^*] + \sum_{j=0}^{2} P[T = t|\hat{v}^* = d(j)] \sum_{i,r} P[I = i, R = r]P[\hat{v}^* = d(j)|I = i, R = r] \\
&\approx& f_{T_{2n-1,m}}(t)P[\hat{v}^* = v^*] + \sum_{j=0}^{2} f_{T_{2n-1,j}}(t) \sum_{i,r} P[I = i, R = r]P[\hat{v}^* = d(j)|I = i, R = r]
\end{aligned}$$

Figure 2 shows an example of the accuracy of the asymtotic distribution for a specific set of parameters.

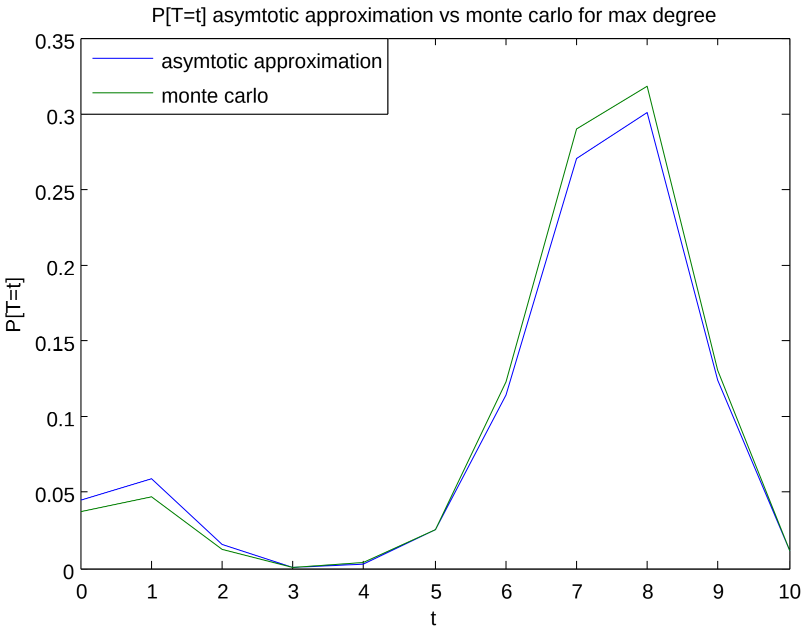

Figure 2: A comparison of a monte carlo simulation and asymtotic distribution of the max degree method with $n = 20$, $m = 10$, $k = 10$, $p = 0.45$, $q = 0.55$, $s_0 = s_1 = 100$.

2.3 Relative Efficiency

With these two methods we need a way to compare their performance. Relative efficiency is one way to do so. Recall that the ratio Nt /NW is a measure of the relative efficiency of the Wilcoxon test versus the t test, where Nt and NW are minimum sample sizes required to achieve some specified power at some specified size. [cite B&D 1977 page 352] [cite Priebe gwmw papers . . . ?]

Since we are in the case for which all signal edges are created equally and all noise edges are created equally, if we constrain all canonical subspace identification methods so that $|\hat{\mathcal{E}}| = k$ for some k, then Priebes Conjecture #1 above implies that comparing the number of signal dimensions recovered for two canonical subspace identification methods allows comparison of classification performance.

Toward that end, for canonical subspace identification method x define

$$T_x(k, s, F_{GY}) = |\mathcal{E} \cap \hat{\mathcal{E}}_x|$$

to be the number of signal dimensions recovered with training sample size s using method x.

Let

$$s_x(t) = \min\{s : E[T_x(k, s, F_{GY})] \geq t\}.$$

The ratio $r(t) = s_{coherent}(t)/s_{agnostic}(t)$ is the relative efficiency.

2.4 Simulations

There do exist situations where the agnostic model outperforms the coherenet model. Notice that in our coherent model, if the wrong vertex is chosen, then $t = 0$. Here is a simulation that shows situations where the agnostic models is better and situations where the coherent model is better. In all simulations $m = 10$, $k = 10$, $t = 8$ and $n = |V|$ ranges from 10 to 150 in intervals of 10. (m is the number of edges with probability q.) For each simulation of $E[T]$, 1000 trials are used.

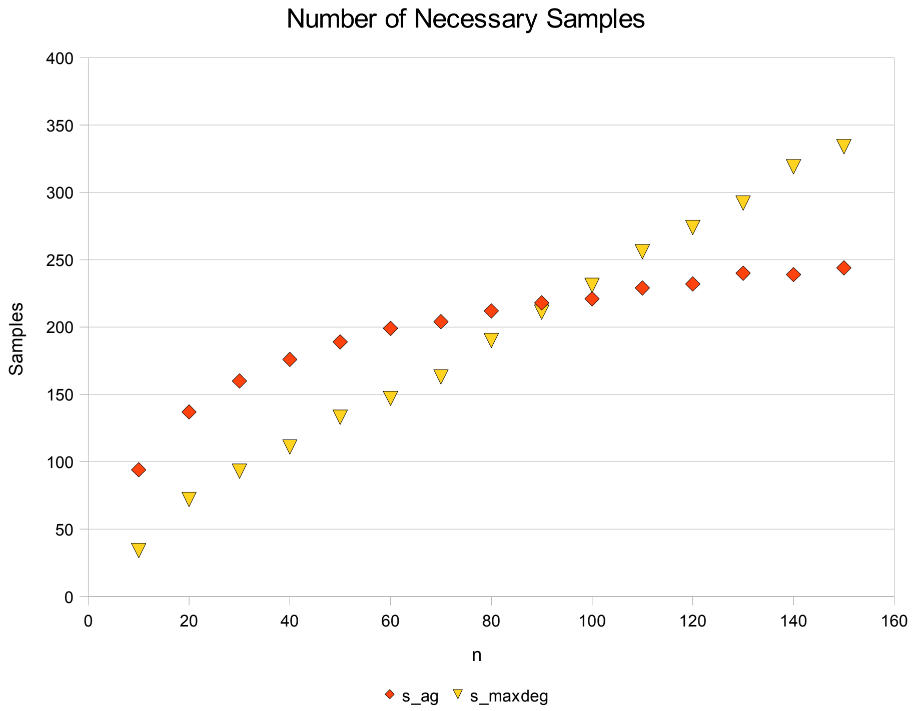

Figure 3: s values of the agnostic and max degree models with $p = 0.1$, $q = 0.2$.

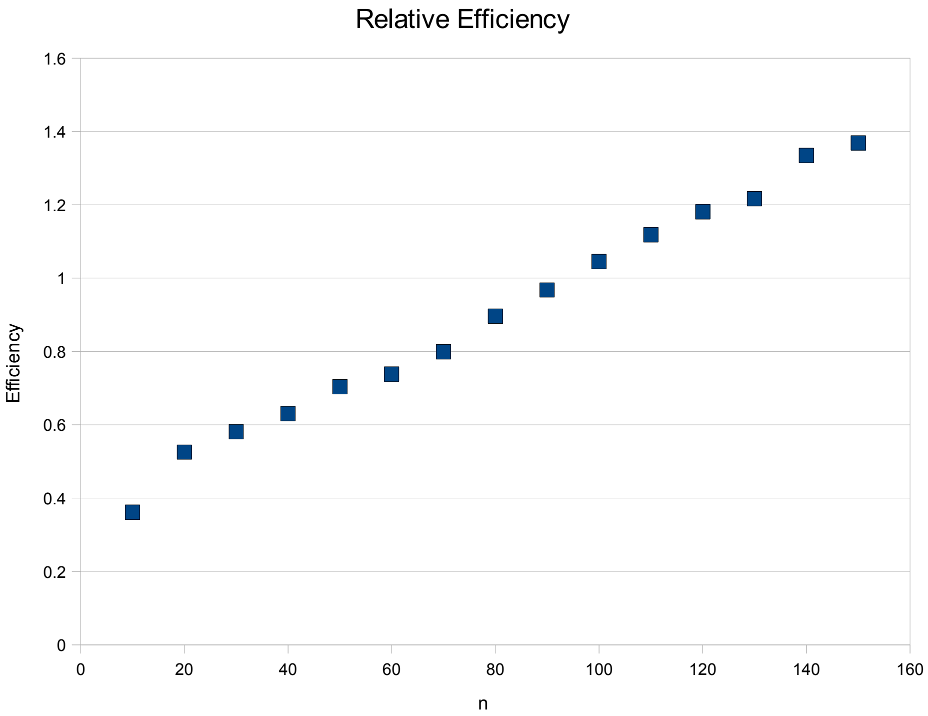

Figure 4: Relative efficiency simulation with $p = 0.1$, $q = 0.2$.

References

[1] H.A. David and H. M. Nagaraja. *Ordered Statistics*. John Wiley & Sons, Inc., third edition edition, 2003.

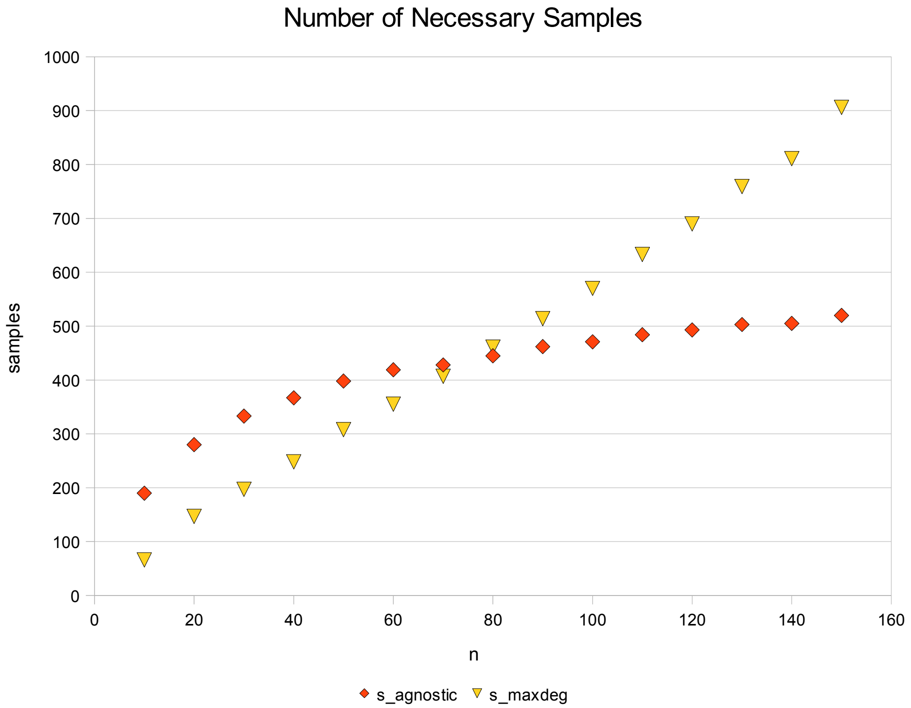

Figure 5: s values of the agnostic and max degree models with $p = 0.45$, $q = 0.55$.

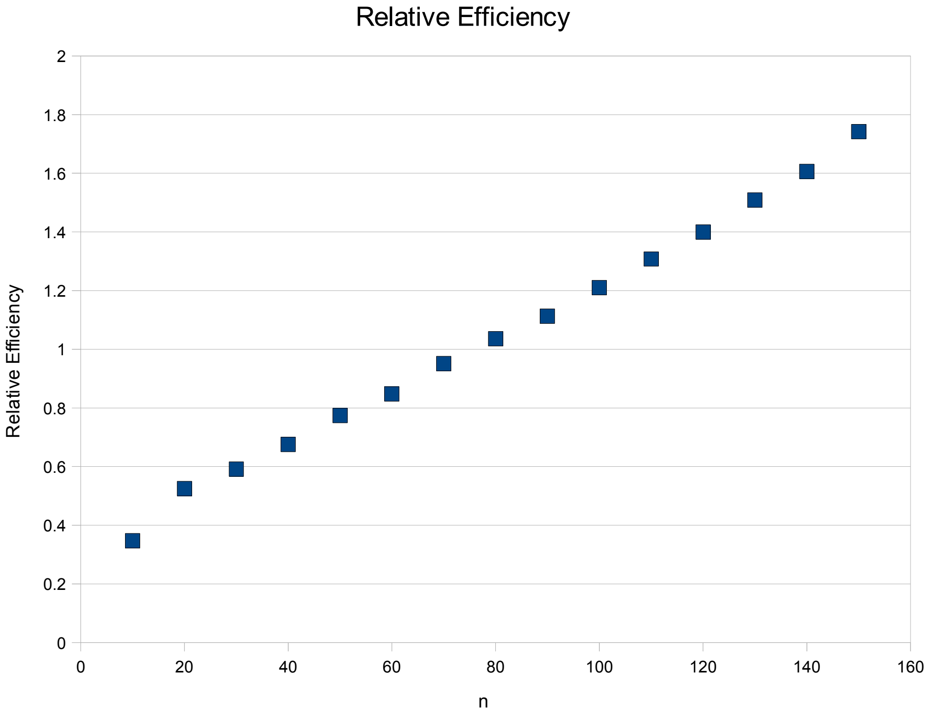

Figure 6: Relative efficiency simulation with $p = 0.45$, $q = 0.55$.

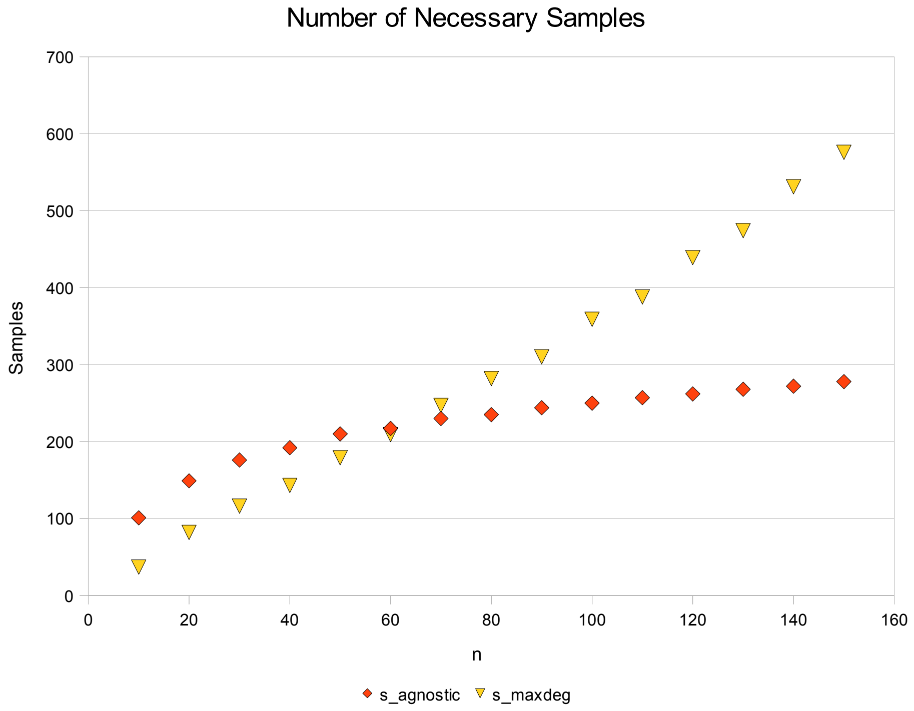

Figure 7: s values of the agnostic and max degree models with $p = 0.8$, $q = 0.9$.

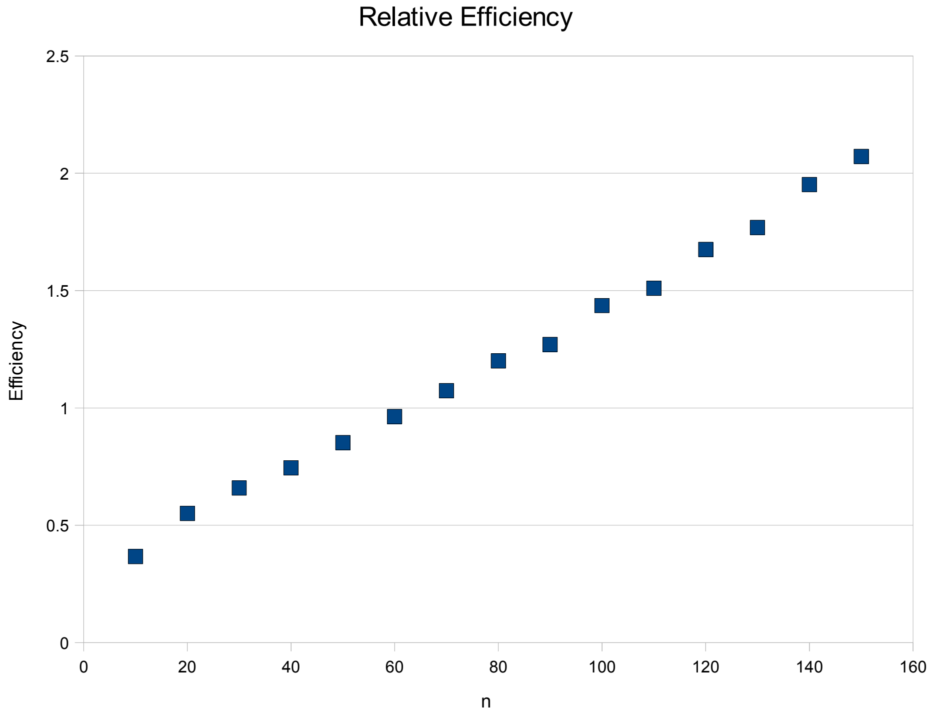

Figure 8: Relative efficiency simulation with $p = 0.8$, $q = 0.9$.