ST 790 Assignment 1

David Elsheimer and Jimmy Hickey

9/1/2020

Instruction

This assignment consists of 4 problems. The assignment is due on **Tuesday, September 1** at 11:59pm EDT. Please submit your assignment electronically through the **Moodle** webpage. The assignment can be done as a group with at most 3 members per group (please include the name of the group members on the front page of the assignment).

Problem 1

Download the political blogs dataset from here. Now find the say, top ten, most "central" or "important" vertices via e.g., closeness centrality, betweeness centrality, Katz centrality, page-rank centrality, and hubs/authority centrality. Does the blogs that you identified similar to that in Table 1 through Table 3 of Adamic and Glance ?

Different Centrality Metric Rankings (Overall, Liberal, & Conservative)

```
g<-read.graph("polblogs.gml",format=c("gml"))
summary(g)
```

```
## IGRAPH 6296513 D--- 1490 19090 --
## + attr: id (v/n), label (v/c), value (v/n), source (v/c)
```

```
labels <- V(g)$label
val <- V(g)$value

closeness <- closeness(g)
betweenness <- betweenness(g)
pg <- page_rank(g)$vector
katz <- alpha_centrality(g, alpha = 0.05)
hub <- hub_score(g)$vector
auth <- authority_score(g)$vector

nodeinfo <- as.data.frame(labels)
nodeinfo <- mutate(nodeinfo, val = val, close=closeness, between=betweenness, pg = pg, katz = katz, hub
```

```
pgrank    <- nodeinfo %>% arrange(desc(close)) %>%  select(labels) %>%head(n=10)

pgranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(pg)) %>%  select(labels) %>%head(n=10)

pgrankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(pg)) %>%  select(labels) %>%head(n=10)


closerank    <- nodeinfo %>% arrange(desc(close)) %>%  select(labels) %>%head(n=10)

closeranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(close)) %>%  select(labels) %>%head(n=10)

closerankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(close)) %>%  select(labels) %>%head(n=10)


betweenrank    <- nodeinfo %>% arrange(desc(between)) %>%  select(labels) %>%head(n=10)

betweenranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(between)) %>%  select(labels) %>%head(n=10)

betweenrankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(between)) %>%  select(labels) %>%head(n=10)

katzrank    <- nodeinfo %>% arrange(desc(katz)) %>%  select(labels) %>%head(n=10)

katzranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(katz)) %>%  select(labels) %>%head(n=10)

katzrankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(katz)) %>%  select(labels) %>%head(n=10)

hubrank    <- nodeinfo %>% arrange(desc(hub)) %>%  select(labels) %>%head(n=10)

hubranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(hub)) %>%  select(labels) %>%head(n=10)

hubrankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(hub)) %>%  select(labels) %>%head(n=10)



authrank    <- nodeinfo %>% arrange(desc(auth)) %>%  select(labels) %>%head(n=10)

authranklib <- nodeinfo %>% filter(val == 0) %>%
  arrange(desc(auth)) %>%  select(labels) %>%head(n=10)

authrankcon <- nodeinfo %>% filter(val == 1) %>%
  arrange(desc(auth)) %>%  select(labels) %>%head(n=10)

rankings1 <- as.data.frame(cbind(pgrank, closerank, betweenrank))
```

```
rankings2 <- as.data.frame(cbind(katzrank, hubrank, authrank))
colnames(rankings1) <- c("Page rank", "Closeness", "Betweeness")
colnames(rankings2) <- c("Katz", "Hub", "Authority")

librankings1 <- as.data.frame(cbind(pgranklib, closeranklib, betweenranklib))

librankings2 <- as.data.frame(cbind(katzranklib, hubranklib, authranklib))
colnames(librankings1) <- c("Page rank", "Closeness", "Betweeness")
colnames(librankings2) <- c("Katz", "Hub", "Authority")

conrankings1 <- as.data.frame(cbind(pgrankcon, closerankcon, betweenrankcon))

conrankings2 <- as.data.frame(cbind(katzrankcon, hubrankcon, authrankcon))
colnames(conrankings1) <- c("Page rank", "Closeness", "Betweeness")
colnames(conrankings2) <- c("Katz", "Hub", "Authority")
```

Overall

	Page rank	Closeness	Betweeness
1	itlookslikethis.blogeasy.com	itlookslikethis.blogeasy.com	blogsforbush.com
2	bushmisunderestimated.blogspot.com	bushmisunderestimated.blogspot.com	atrios.blogspot.com
3	etherealgirl.blogspot.com	etherealgirl.blogspot.com	instapundit.com
4	michaelphillips.blogspot.com	michaelphillips.blogspot.com	dailykos.com
5	lennonreport.blogspot.com	lennonreport.blogspot.com	newleftblogs.blogspot.com
6	isdl.blogspot.com	isdl.blogspot.com	madkane.com/notable.html
7	isdl.weblogs.us	isdl.weblogs.us	wizbangblog.com
8	janm.blogspot.com	janm.blogspot.com	lashawnbarber.com
9	nerofiddled.blogspot.com	nerofiddled.blogspot.com	hughhewitt.com
10	saveoursenate.blogspot.com	saveoursenate.blogspot.com	washingtonmonthly.com

	Katz	Hub	Authority
1	rooksrant.com	politicalstrategy.org	dailykos.com
2	blogsagainsthillary.com	madkane.com/notable.html	talkingpointsmemo.com
3	burntorangereport.com	liberaloasis.com	atrios.blogspot.com
4	collegedems.com/blog	stagefour.typepad.com/commonprejudice	washingtonmonthly.com
5	gregpalast.com	bodyandsoul.typepad.com	talkleft.com
6	bopnews.com	corrente.blogspot.com	instapundit.com
7	rudepundit.blogspot.com	atrios.blogspot.com/	juancole.com
8	iddybud.blogspot.com	tbogg.blogspot.com	yglesias.typepad.com/matthew
9	shakespearessister.blogspot.com	newleftblogs.blogspot.com	pandagon.net
10	home.earthlink.net/~fsrhine	atrios.blogspot.com	digbysblog.blogspot.com

Liberal

	Page rank	Closeness	Betweeness
1	dailykos.com	itlookslikethis.blogeasy.com	atrios.blogspot.com
2	atrios.blogspot.com	bushmisunderestimated.blogspot.com	dailykos.com
3	talkingpointsmemo.com	etherealgirl.blogspot.com	newleftblogs.blogspot.com
4	washingtonmonthly.com	michaelphillips.blogspot.com	madkane.com/notable.html
5	juancole.com	lennonreport.blogspot.com	washingtonmonthly.com
6	prospect.org/weblog	janm.blogspot.com	liberaloasis.com
7	digbysblog.blogspot.com	nerofiddled.blogspot.com	digbysblog.blogspot.com
8	talkleft.com	saveoursenate.blogspot.com	robschumacher.blogspot.com
9	yglesias.typepad.com/matthew	dashnier.blogspot.com	nomoremister.blogspot.com
10	jameswolcott.com	politicalmonitor.us/blog	xnerg.blogspot.com

	Katz	Hub	Authority
1	rooksrant.com	politicalstrategy.org	dailykos.com
2	burntorangereport.com	madkane.com/notable.html	talkingpointsmemo.com
3	collegedems.com/blog	liberaloasis.com	atrios.blogspot.com
4	gregpalast.com	stagefour.typepad.com/commonprejudice	washingtonmonthly.com
5	bopnews.com	bodyandsoul.typepad.com	talkleft.com
6	rudepundit.blogspot.com	corrente.blogspot.com	juancole.com
7	iddybud.blogspot.com	atrios.blogspot.com/	yglesias.typepad.com/matthew
8	shakespearessister.blogspot.com	tbogg.blogspot.com	pandagon.net
9	home.earthlink.net/~fsrhine	newleftblogs.blogspot.com	digbysblog.blogspot.com
10	mahablog.com	atrios.blogspot.com	prospect.org/weblog

Conservative

	Page rank	Closeness	Betweeness
1	instapundit.com	isdl.blogspot.com	blogsforbush.com
2	blogsforbush.com	isdl.weblogs.us	instapundit.com
3	michellemalkin.com	streetlog.typepad.com	wizbangblog.com
4	drudgereport.com	rednyc.blogspot.com	lashawnbarber.com
5	powerlineblog.com	r2korn.blogdrive.com	hughhewitt.com
6	andrewsullivan.com	thelastword.blogdrive.com	gevkaffeegal.typepad.com/the_alliance
7	littlegreenfootballs.com/weblog	calirep.blogspot.com	michellemalkin.com
8	vodkapundit.com	smashingzero.blogspot.com	truthlaidbear.com
9	rightwingnews.com	lonewacko.com	littlegreenfootballs.com/weblog
10	volokh.com	nicoladellarciprete.blogs.com	evangelicaloutpost.com

	Katz	Hub	Authority
1	blogsagainsthillary.com	instapundit.com	instapundit.com
2	etalkinghead.com	dalythoughts.com	powerlineblog.com
3	nickiegoomba.blogspot.com	acertainslantoflight.blogspot.com	andrewsullivan.com
4	swiftreport.blogs.com	incite1.blogspot.com	truthlaidbear.com
5	stmonk.blogspot.com	thomasgalvin.blogspot.com	michellemalkin.com
6	poeticvalues.blogspot.com	truthprobe.blogspot.com	drudgereport.com
7	thejessefactor.blogspot.com	blogsofwar.com	nationalreview.com/thecorner
8	conservativepunk.com	scha-den-freu-de.blogspot.com	littlegreenfootballs.com/weblog
9	zebrax.blogs.com	vodkapundit.com	hughhewitt.com
10	dizzy-girl.net/index.php	cayankee.blogs.com	volokh.com

Using the page rank metric, the top 10 blocks for liberal and conservative are similar for to those in tables 1 and 2. The top 10 vary based on the metric chosen, but several other top 10 rankings are similar as well. For example, the authority and betweeness centrality metrics are fair similar for conservative blogs as well as liberal blogs. For the overall rankings (table 3 in the paper), the metrics used to rank blogs are unclear. There does appear to be some consensus between the different centrality metrics, however.

Problem 2

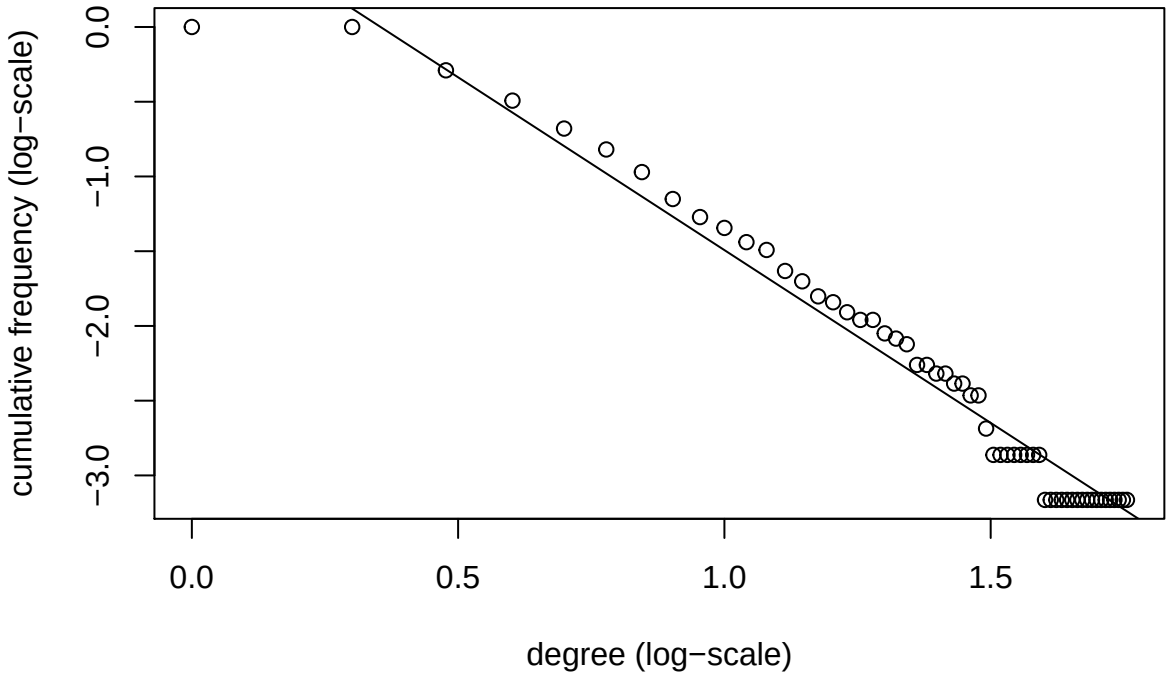
Skim through this article by Clauset, Shalizi, and Newman. Now download the yeast metabolic network dataset from here. Fit a power-law to the degree distribution for this dataset. Do you think the degree distribution for this network follows a power-law ?

```
library(igraph)
g2 <- read_graph("nodes.txt")
cg <- components(g2)
lcc <- decompose(g2, min.vertices = max(cg$csize))[[1]];

deg_dist <- degree_distribution(lcc, cumulative = TRUE)
mod <- lm(log(deg_dist,10) ~ log(1:length(deg_dist),10))
mod$coefficients ##``Estimate'' of exponent in power-law
```

```
##                 (Intercept) log(1:length(deg_dist), 10)
##                   0.8195035                  -2.3120220
```

```
plot(log(1:length(deg_dist),10),log(deg_dist,10),
xlab = "degree (log-scale)", ylab = "cumulative frequency (log-scale)",abline(mod))
```



Based on the above comparison of the data to the fitted power law, it appears that the degree distribution does indeed follow a power law.

Problem 3

Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables. The sample variance of the X_i can be written as a U-statistics, i.e.,

$$s^2 = \frac{2}{n(n-1)} \sum_{i<j} \frac{1}{2}(X_i - X_j)^2.$$

Suppose now the X_i are i.i.d Bernoulli random variables with probability of success $p \in (0, 1)$. Derive a non-degenerate limiting distribution for s^2. What happens to this limiting distribution when $p = 1/2$?

$$
\begin{aligned}
s^2 = \frac{2}{n(n-1)} \sum_{i<j} \frac{1}{2}(X_i - X_j)^2 &= \binom{n}{2}^{-1} \sum_{i<j} \frac{1}{2}(X_i - X_j)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
(n-1)s^2 &= \sum_{i=1}^{n} ((X_i - p) - (\bar{X} - p))^2 \\
&= \sum_{i=1}^{n} (X_i - p)^2 - 2\sum_{i=1}^{n} (X_i - p)(\bar{X} - p) + \sum_{i=1}^{n} (\bar{X} - p)^2 \\
&= \sum_{i=1}^{n} (X_i - p)^2 - n(\bar{X} - p)^2 \\
\sqrt{n}(s^2 - p(1-p)) &= \frac{\sqrt{n}}{n-1} \sum_{i=1}^{n} (X_i - p)^2 - \sqrt{n}p(1-p) - \frac{\sqrt{n}}{n-1} n(\bar{X} - p)^2 \\
&= \frac{\sqrt{n}}{n-1} \sum_{i=1}^{n} (X_i - p)^2 - \sqrt{n}\frac{n-1}{n-1}p(1-p) - \frac{n}{n-1}\sqrt{n}(\bar{X} - p)^2 \\
&= \frac{n\sqrt{n}}{n-1} \frac{1}{n} \sum_{i=1}^{n} (X_i - p)^2 - \sqrt{n}\frac{n-1}{n-1}p(1-p) - \frac{n}{n-1}\sqrt{n}(\bar{X} - p)^2 \\
&= \frac{n}{n-1} \left(\sqrt{n}(\frac{1}{n} \sum_{i=1}^{n} (X_i - p)^2 - p(1-p)) \right) + \frac{\sqrt{n}}{n-1} p(1-p) - \frac{n}{n-1}\sqrt{n}(\bar{X} - p)^2
\end{aligned}
$$

Note as $n \to \infty$, $\frac{n}{n-1} \to 1$, $\frac{\sqrt{n}}{n-1} \to 0$. Furthermore, note that $\sqrt{n}(\bar{X} - p)^2 = \sqrt{n}(\bar{X} - p)(\bar{X} - p)$. $(\bar{X} - p) \xrightarrow{p} 0$, $\sqrt{n}(\bar{X} - p) \xrightarrow{d} N(0, p(1-p))$ by the Central Limit Theorem. As such, by Slutsky's, $\sqrt{n}(\bar{X} - p)^2 \xrightarrow{d} 0$.

Thus the problem simplies to finding the limiting distribution of $\frac{1}{n} \sum_{i=1}^{n} (X_i - p)^2$ using the central limit theorem. $E((X_i - p)^2) = E(X_i^2 - 2pX_i + p^2) = p^2 + p(1-p) - 2p^2 + p^2 = p(1-p)$. $V(X_i - p)^2 = E((X_i - p)^4) - (p(1-p)^2) = \mu_4 - (p(1-p))^2$. For a bernoulli distribution, the 4th central moment is $p(1-p)(1-3p(1-p))$.

Thus as $n \to \infty$, $\sqrt{n}(s^2 - p(1-p)) \equiv \sqrt{n}(\frac{1}{n} \sum_{i=1}^{n} (X_i - p)^2 - p(1-p)) \xrightarrow{d} N(0, p(1-p)(1-3p(1-p)) - (p(1-p))^2)$ by the Central Limit Theorem.

When $p = \frac{1}{2}$, $p(1-p)(1-3p(1-p)) - (p(1-p))^2) = 0$, resulting in a degenerate point mass.

Problem 4

We now consider the problem of goodness-of-fit test for Erdos-Renyi graphs. More specifically, let $\mathcal{G} = \mathcal{G}(n, m, p, q)$ be a random graph model defined as follows. A graph G is an instance of $\mathcal{G}$ if G is an undirected graph on n vertices with adjacency matrix $\mathbf{A} = (a_{ij})$ of the form

$$
\begin{aligned}
a_{ij} &\sim \text{Bernoulli}(q), \quad \text{if } i \in \mathcal{S} \text{ and } j \in \mathcal{S} \\
a_{ij} &\sim \text{Bernoulli}(p), \quad \text{otherwise.}
\end{aligned}
$$

Here $\mathcal{S} \subset \{1, 2, \ldots, n\}$ is a subset of vertices with $|\mathcal{S}| = m$ and we have assumed that $q > p \in (0, 1)$. Note that $m = 0$ corresponds to the normal Erdos-Renyi graph and $m > 0$ corresponds to a random graph model where there is a subset $\mathcal{S}$ of m vertices with larger communication probability q among themselves.

Given a graph $G \sim \mathcal{G}(n, m, p, q)$, we are interested in testing the null hypothesis $\mathbb{H}_0 \colon |\mathcal{S}| = m = 0$ against the alternative hypothesis that $\mathbb{H}_A \colon |\mathcal{S}| = m > 0$.

For this problem, we will use two test statistics. The first test statistic is Δ, the maximum degree and the second test statistic is τ, the number of triangles. More specifically, given a graph $G \sim \mathcal{G}(n, m, p, q)$, and suppose we are using Δ as the test statistic. Then we will reject the null hypothesis if Δ exceeds some threshold c.

Perform a simulation study to determine the power of the test procedures when we use (1) Δ as the test statistic and (2) when we use τ as the test statistic. To reduce computation time, set $n = 1000$, $p = 0.4$, and $q = 0.6$ and let $m \in \{10, 25, 50, 100\}$. Report a table for the power of the test statistics as a function of m.

Hint To generate a graph from the $\mathcal{G}(n, m, p, q)$ when $m > 0$, you can use the following igraph chunk.

```
delta_power = function(p, q, n, m, alpha, reps)
{
  # Extreme Value Distributions package for Gumbel distribution
  require(evd)
  require(igraph)
  pvals = rep(0,reps)


  for (i in 1:reps)
  {
    B <- matrix(c(p,p,p,q),nrow = 2)
    A <- sbm.game(n = n, pref.matrix = B, block.sizes = c(n-m,m))
    delta = max(degree(A))

    # naive estimate of edge probability
    num_edges = ecount(A)
    phat =  num_edges / choose(n, 2)

    y = -2 * log(n) + (sqrt(2) *(-n * phat + delta) * log(n) ) / (sqrt(n * phat *(1-phat) * log(n))) + 

    pvals[i] =1- pgumbel(y, 0 , 1)
  }

  return(sum(pvals<=alpha)/length(pvals))
}

n_reps = 100
p <- 0.4
q <- 0.6
n <- 1000
alpha = 0.05

delta_10 = delta_power(p,q,n, m= 10, alpha, n_reps)
```

```
## Warning: package 'evd' was built under R version 3.6.3
```

```
delta_25 = delta_power(p,q,n, m= 25, alpha, n_reps)
delta_50 = delta_power(p,q,n, m= 50, alpha, n_reps)
delta_100 = delta_power(p,q,n, m= 100, alpha, n_reps)

cat("delta_10\t", delta_10, "\ndelta_25\t", delta_25, "\ndelta_50\t", delta_50, "\ndelta_100\t", delta_
```

```
## delta_10  0.07
## delta_25  0.06
## delta_50  0.03
## delta_100     0.2
```

```
### Triangles

tau_power = function(p, q, n, m, alpha, reps)
{
  require(igraph)

  pvals = rep(1, reps)
  for (i in 1:reps)
  {
    B <- matrix(c(p,p,p,q),nrow = 2)
    A <- sbm.game(n = n, pref.matrix = B, block.sizes = c(n-m,m))

    # naive estimate of edge probability
    num_edges = ecount(A)
    phat =  num_edges / choose(n, 2)

    tau = sum(count_triangles(A))/3

    tau_test = (tau - choose(n,3)*phat^3) / ((n - 2) * p^2 * sqrt( choose(n, 2) * phat * (1-phat)))

    pvals[i] = 1 - pnorm(tau_test, 0, 1)
  }

  return(sum(pvals<=alpha)/length(pvals))
}

tau_10 =tau_power(p,q,n, m= 10, alpha, n_reps)
tau_25 =tau_power(p,q,n, m= 25, alpha, n_reps)
tau_50 =tau_power(p,q,n, m= 50, alpha, n_reps)
tau_100 =tau_power(p,q,n, m= 100, alpha, n_reps)
tau_200 =tau_power(p,q,n, m= 200, alpha, n_reps)
tau_400 =tau_power(p,q,n, m= 400, alpha, n_reps)

cat("tau_10\t", tau_10, "\ntau_25\t", tau_25, "\ntau_50\t", tau_50, "\ntau_100\t", tau_100,
    "\ntau_200\t", tau_200,"\ntau_400\t", tau_400)
```

```
## tau_10    0
## tau_25    0
## tau_50    0
## tau_100   0
## tau_200   0
## tau_400   1
```

As seen for both cases, as m increases, power increases. However, the test statistic for the number of triangles is smaller than the test statistic for max degree, and thus for meaningful results, greater levels of m must be considered for that test.