Graphs and Networks	**out:** October 5, 2010

Problem Set 2

Lecturer: Daniel A. Spielman	**due:** October 19, 2010

1 Introduction

There are 3 sections to this problem set. The first contains problems for everyone. The second contains problems for those who are taking the "theoretical track". The third contains experimental problems for those who are taking the "experimental track".

2 Homework Policy

You may discuss the problems with other students. But, you must write your solutions independently, drawing on your own understanding. You should cite any sources that you use on the problem sets other than the textbook, TA and instructor. This means that you should list your collaborators.

If you are doing the experimental problems, it goes without saying that you should write your own code.

You **may not** search the web for solutions to similar problems given out in other classes. If you think this policy needs any clarification, please let me know.

3 Clarifications

1. In line a of problem 3, I previously write $\boldsymbol{w}_1$ when I meant $\boldsymbol{u}_1$. This is fixed in this version.
2. In the experimental problems 3-5, you wind up with a function on the vertices. Let's call it f. This function determines a sequence of sets S_k, where each S_k is the set of k vertices for which f is lowest (or highest). You should then find the k for which $\phi(S_k)$ is lowest. This will, hopefully, give a set of low conductance.
3. In the experimental problems 4 and 5, when I say "choose 10 vertices at random to start", I mean to choose 1 vertex and random, run the process, and then repeat this 10 times (with different starting vertices).
4. In the experimental problems, I wrote "$\max(n, \lceil 10,000/\sqrt{n} \rceil)$" when I meant to write "min".
5. I also re-wrote what you are supposed to prove in parts 4c and 4e of the theoretical problems.

4 Problems for all

1. Prove parts a and b of Lemma 2.1 from Lecture 10.

5 Theoretical Problems

2. Let $\boldsymbol{W}$ be the *ordinary* walk matrix of a connected graph. As in class, for any initial probability distribution $\boldsymbol{p}_0$, let $\boldsymbol{p}_t = \boldsymbol{W}^t \boldsymbol{p}_0$. Prove that as t grows large
$$\boldsymbol{p}_t + \boldsymbol{p}_{t+1}$$
approaches a limit.

3a. Let $\widehat{\boldsymbol{W}}$ be the *lazy* walk matrix of a connected graph. Let $\boldsymbol{p}$ be a probability vector and let α be a number such that
$$\boldsymbol{p}(u) \leq \alpha d(u),$$
for all vertices u. Prove that
$$(\widehat{\boldsymbol{W}} \boldsymbol{p})(u) \leq \alpha d(u),$$
for all vertices u.

3b. Let $\widehat{\boldsymbol{W}}$ be the *lazy* walk matrix of a connected graph. Let S be any set of vertices and let $\boldsymbol{p}_0 = \boldsymbol{\pi}_S$ be the restriction of the steady-state distribution of S. Prove that for every $t \geq 0$,
$$\boldsymbol{p}_t(S) \geq 1 - t\frac{|\partial(S)|}{d(S)}.$$

4. In this problem, we study the directed 1-out preferential attachment model from Lecture 7. As in the second part of the lecture, define $X_d(t)$ to be the number of nodes of in-degree d at time t. At the end of the lecture, I said that we could prove that this was tightly concentrated for small d by using a Martingale argument. Such an argument reasons about variables Z_d^k, which we define to be the expected number of in-degree d nodes at time n, given the choices of edges from the first k nodes.

   Let me make that more concrete. For a node i, let $e(i)$ be the destination of the edge leaving i. We define
$$Z_d^k = \mathbf{E}_{e(k+1),\ldots,e(n)} \left[X_d^n \middle| e(1), ..., e(k) \right].$$

   Note that Z_d^k is a function of $e(1), \ldots, e(k)$. Thus, the expected number of degree d nodes at the end is Z_d^0, and
$$\mathbf{E}_{e(1),\ldots,e(k)} \left[Z_d^k \right] = Z_d^0.$$

   The final number of degree d nodes is Z_d^n. One can use Azuma's inequality to prove that these are highly concentrated if one can prove that for all k and all $e(1), \ldots, e(k), e(k+1)$
$$\left| Z_d^k \left(e(1), \ldots, e(k)\right) - Z_d^{k+1} \left(e(1), \ldots, e(k+1)\right) \right| \leq c.$$

for some constant c. In this problem, we will see two different ways of proving this. In the first part, we just worry about the nodes of in-degree 0.

a. Given that node j has in-degree 0 after step t, what is the chance it finishes with in-degree 0?

b. Assume that after step t there are x nodes with in-degree 0. What is the expected number of nodes of in-degree 0 at the end?

c. Use parts a and b to prove that for all k and for all $e(1), \ldots, e(k), e(k+1)$

$$\left| Z_0^k\left(e(1), \ldots, e(k)\right) - Z_0^{k+1}\left(e(1), \ldots, e(k+1)\right) \right| \leq 1.$$

d. In class, we described the process by saying that, for each vertex t, with probability p the destination $e(t)$ is chosen uniformly from $\{1, \ldots, t-1\}$ and with probability $1-p$ it is chosen according to degree. Let's re-phrase this. First, a vertex $j \in \{1, \ldots, t-1\}$ should be chosen uniformly at random. Then, with probability p we set $e(t) = j$ and with probability $1-p$ we set $e(t) = e(j)$. Prove that if only one of these choices is made differently during the process then the degrees of at most two nodes can differ at the end.

   Given this, one can prove concentration of all the variables $X_d(n)$ by using Talagrand's inequality. Or, we can go through the Martingale route:

e. Use the result of part d to prove that for all d, k, and $e(1), \ldots, e(k+1)$,

$$\left| Z_d^k\left(e(1), \ldots, e(k)\right) - Z_d^{k+1}\left(e(1), \ldots, e(k+1)\right) \right| \leq 2.$$

6 Experiments in Conductance

In this homework, we will look for sets of small conductance in the graph.

2. Use the metadata associated with your vertices or edges to construct *interesting* sets of vertices in your graph. Measure the conductance of these sets. Report your results in some reasonable way.

   If you cannot use your metadata to construct interesting sets of vertices, then get a new graph.

3. Compute a good approximation to $\boldsymbol{v}_2$ and λ_2 for the lazy random walk on your graph. Report the value of λ_2 that you find.

   The easiest way to compute $\boldsymbol{v}_2$ is to first form the symmetric matrix $\widehat{\boldsymbol{M}}$. Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ denote the eigenvectors of $\widehat{\boldsymbol{M}}$ corresponding to the first and second-largest eigenvalues. You can compute $\boldsymbol{u}_1$ explicitly. Now, choose a uniformly random vector $\boldsymbol{r}$ (ideally with each component being an independent Gaussian). If you don't know how to compute eigenvectors, you can get a good approximation to $\boldsymbol{u}_2$ by setting $\boldsymbol{x} = \boldsymbol{r}$ and then iterating the following three operations:

a. Set $\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{u}_1(\boldsymbol{u}_1^T\boldsymbol{x})$ (orthogonalize $\boldsymbol{x}$ with respect to $\boldsymbol{u}_1$).

b. Set $\boldsymbol{x} = \boldsymbol{x}/\|\boldsymbol{x}\|$.

c. Set $\boldsymbol{x} = \widehat{\boldsymbol{M}}\boldsymbol{x}$.

Now, you just need to decide when you can stop. Given $\boldsymbol{x}$, an approximation to $\boldsymbol{u}_2$, you can estimate λ_2 by

$$\lambda = \max_i \frac{(\widehat{\boldsymbol{M}}\boldsymbol{x})(i)}{\boldsymbol{x}(i)}.$$

Set $\mu = 1 - \lambda$. You are close enough to convergence if for every i

$$\left|(\widehat{\boldsymbol{M}}\boldsymbol{x})(i) - \lambda\boldsymbol{x}(i)\right| \leq \mu\boldsymbol{x}(i).$$

Now that you've got an approximation to $\boldsymbol{u}_2$, use it to derive an approximation to $\boldsymbol{v}_2$ and use that to find a set S of low conductance. What is the conductance of the set that you find? How does it compare to what you would expect from Cheeger's inequality.

4. Now, let's try to find sets of low conductance by computing the distributions of random walks. If your graph has more than $10,000$ vertices, do this by choosing 10 vertices at random to start. If your graph has fewer than $10,000$ vertices, do this by choosing at least $\min(n, \lceil 10,000/\sqrt{n}\rceil)$ random vertices to start. For each starting distribution, evolve the random walk until $\|\boldsymbol{p}_t - \boldsymbol{\pi}\| \leq \|\boldsymbol{\pi}\|/1000$. For each distribution encountered at every step of each walk, compute the set of lowest conductance that you can find from the distribution (i.e. considering the k vertices maximizing $\boldsymbol{p}_t(i)/d(i)$). For each starting point, keep track of the least conductance that you can find.

   If you just did 10 runs, then report all these conductances. If you did more, report the best few conductances and plot the rest in some reasonable way.

5. Do the same thing as in step 4, but starting from 10 random points in the set S that you found in part 2 that had the least conductance.