T-79.7003: Graphs and Networks	**Out: Sept. 27th, 2013**
Homework 1	
Lecturer: Charalampos E. Tsourakakis	*In: Oct. 11th, 2013*

Comments

If you write more than 100 points, they will count as bonus. You are all required to solve 1.2(B), 1.3(A) through (E) and 1.3(H).

1.1 Probabilistic inequalities [30 points]

(A) Cauchy-Schwartz inequality [10 points] Prove the Cauchy-Schwartz inequality for random variables X, Y

$$\boxed{|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.}$$

(B) Bonferonni Inequalities [10 points] Let $E_1, E_2, \ldots, E_n$ be events in a sample space. We have been using the union bound a lot in our class:

$$\mathbf{Pr}[E_1 \cup \ldots \cup E_n] \leq \sum_{i=1}^{n} \mathbf{Pr}[E_i].$$

In this exercise you will prove a more general result. Define

$$S_1 = \sum_{i=1}^{n} \mathbf{Pr}[E_i]$$
$$S_2 = \sum_{i<j} \mathbf{Pr}[E_i \cap E_j]$$

and for $2 < k \leq n$,

$$S_k = \sum_{(i_1,\ldots,i_k)} \mathbf{Pr}[E_{i_1} \cap \ldots \cap E_{i_k}],$$

where the summation is taken over all ordered k-tuples of distinct integers.

<u>Prove</u> for *odd* k, $1 \leq k \leq n$

$$\boxed{\mathbf{Pr}\left[E_1 \cup \ldots \cup E_n\right] \leq \sum_{j=1}^{k} (-1)^{j+1} S_j.}$$

and for *even* k, $2 \leq k \leq n$

$$\boxed{\mathbf{Pr}\left[E_1 \cup \ldots \cup E_n\right] \geq \sum_{j=1}^{k} (-1)^{j+1} S_j.}$$

(C) [10 points] Let $\mathcal{A} = \{A_1, \ldots, A_m\}$ be a collection of events in a probability space. Let $\mu = \sum_{i=1}^{m} \mathbf{Pr}\left[A_i\right]$ be the expected number of events from $\mathcal{A}$ that occur. Given a fixed integer l, let Q be the event that some set of l independent events from $\mathcal{A}$ occur. In other words, Q is the event that, among the events in $\mathcal{A}$ that occur, there are l that are mutually independent. Show that

$$\boxed{\mathbf{Pr}\left[Q\right] \leq \frac{\mu^l}{l!}.}$$

1.2 Erdös-Rényi graphs [55 points]

(A) Practicing the first moment method [5 points] Let $G \sim G(n,p)$ where $p = o(n^{-3/2})$. Prove that G consists of isolated vertices and independent edges.

(B) Cycles in $G(n,p)$ [30 points] Prove that the threshold for the emergence of cycles in $G(n,p)$ is $p^* = \frac{1}{n}$.

(C) Perfect matchings in random bipartite graphs $B(n,n,p)$ [20 points] Let $p = \frac{\log n + c}{n}$ where c is a constant. Let G be a random subgraph of the complete bipartite graph $K_{n,n}$. given by taking each edge with probability p, where choices are made independently. Show that

$$\mathbf{Pr}\left[G \text{ has a perfect matching}\right] \rightarrow e^{-2e^{-c}}$$

as $n \rightarrow +\infty$.

Hints: (a) Use the Bonferroni inequalities to "sandwich" the probability of the event "no vertex is isolated". [10 points] (b) Then, prove that the main reason why there can be no perfect matching in G are isolated vertices. In other words, show that the probability that Hall's theorem is violated for any other reason is $o(1)$. *[10 points]*

1.3 Empirical Properties of Networks [65 points]

In this problem you will study empirically various properties of networks[1]. First, download the following graphs[2]

1. Amazon product co-purchasing network from March 2 2003 from
`http://www.cise.ufl.edu/research/sparse/matrices/SNAP/amazon0302.html`
2. Arxiv High Energy Physics paper citation network from
`http://www.cise.ufl.edu/research/sparse/matrices/SNAP/cit-HepPh.html`
3. Road network of Pennsylvania from
`http://www.cise.ufl.edu/research/sparse/matrices/SNAP/roadNet-PA.html`
4. Web graph of Notre Dame from
`http://www.cise.ufl.edu/research/sparse/matrices/SNAP/web-NotreDame.html`
5. Gnutella peer to peer network from August 9 2002 from
`http://www.cise.ufl.edu/research/sparse/matrices/SNAP/p2p-Gnutella09.html`.

You may use your favorite programming language to code up the following tasks. You may re-use existing software (actually, you should). Check the Web page under the Resources tab to find links to useful packages.

(A) [2 points] For each graph: if it is directed, make it undirected, by ignoring the direction of each edge. Remove multiple edges and self-loops.

(B) [8 points] For each graph:

- Report the number of vertices and edges. Compute the average degree and the variance of the degree distribution.
- Generate the following frequenty plot: the x-axis will correspond to degrees and the y-axis to frequencies. The function you will plot is $f(x) = \#$vertices with degree x. Re-plot the same function in log-log scale.
- Use the code available at `http://tuvalu.santafe.edu/~aaronc/powerlaws/` to fit a power-law distribution to the degree sequence of the graph. Report the output of the *plfit* function.

(C) [10 points] Plot a histogram of the sizes of the connected components of each graph.

(D) [10 points] For each graph, pick any vertex v in the connected component of the largest order. Report the id of the vertex you chose and compute for each $k = 1, 2, ..,\quad f(k) = \#$ vertices at distance k from v. Plot $f(k)$ versus k.

(E) [5 points] For each graph compute the diameter of the largest connected component.

[1] *Send me your code by e-mail.*
[2] The files are .mat. If you are not using MATLAB you can download the same graphs in different format from `http://snap.stanford.edu/data/`.

(F) [10 points] For each graph:

1. Compute for each vertex v in how many K_3s it participates in.
2. Compute the local clustering coefficients and plot their distribution.
3. Let k=degree, $f(k)$ =average number of triangles over all vertices of degree k. Plot $f(k)$ versus k in log-log scale, including error bars for the variance. Fit a least squares line and report the slope.
4. How can you use the previous answer to find outliers in a network?

(G) [5 points] For each graph report the top-20 eigenvalues of the adjacency matrix.

(H) [5 points] For each of the five (5) graphs, generate a random binomial graph on the same number of vertices n_i, where n_i is the number of vertices in G_i, $i = 1, \ldots, 5$ with $p = \frac{2 \log n_i}{n_i}$. Answer questions (A) through (G) for these graphs.

(I) [10 points] Make a high-level evaluation of your findings. For instance, how different is the road network from the Web graph? Also, compare your findings between real-world networks and random binomial graphs.