Computational Systems Biology **Fall Semester, 2007/8**

Assignment 1

The exercise should be submitted by 20/3/08 to Nir Yosef's mailbox in Schreiber.

1 Scale free networks (10 pts)

A degree distribution $P(\cdot)$ is scale free if there exists a function g such that for all a, $P(a \cdot k) = g(a) \cdot P(k)$. Prove that a degree distribution is scale free if and only if it is power-law. You can assume that both $P(\cdot)$ and $g(\cdot)$ are differentiable.

2 Estimating the clustering coefficient (30 pts)

For a given degree sequence, let Φ be the collection of all graphs on the same vertex set V with this degree sequence. Let $n = |V|$ and let m be the number of edges in a graph from Φ.

a. Prove that the chance of having an edge between u and v in a graph from Φ is approximately $\frac{d(u)d(v)}{2m}$, where $d(v)$ is the degree of v.

b. Let $P(\cdot)$ be the degree distribution of graphs in Φ (as defined by the degree sequence). Use your result from item (a) to prove that the clustering coefficient of such graphs is: $C = \frac{M_1}{n} \cdot (\frac{M_2 - M_1}{M_1^2})^2$, where M_1, M_2 are the first and second moments of $P(\cdot)$.

For the rest of the question, assume that $P(\cdot)$ is scale free, $P(k) \propto k^{-c}$.

c. Show that the maximum degree in a graph from Φ is roughly $n^{\frac{1}{c-1}}$, by assuming that there exists a single vertex with this maximum degree.

d. Conclude that for power law distributions with $2 < c < 3$, the clustering coefficient can be estimated as $n^{\frac{3c-7}{c-1}}$.

Hint: Estimate M_2 using the largest term in the sum defining it.

3 K-means clustering (10 pts)

Applying the k-means algorithm on a given set $S \subseteq R^n$ produces a partition of the set into k distinct clusters $\{T_1 \dots T_k\}$, $T_i \subseteq S$.

Given such a cluster $T = \{x_1 \dots x_m\}$ let c_T be a convex sum of its elements, i.e., $c_T = \sum_{j=1}^{m} \lambda_j \cdot x_j$ such that $\sum_{j=1}^{m} \lambda_j = 1$, and $\lambda_j \geq 0$. Prove that the closest centroid (out of the centers of all the clusters in the partition) to c_T is the centroid of T.
What does this result imply on the applicability of k-means?

4 Biclustering (20 pts)

In this question you will analyze the complexity of graph problems related to biclustering.

a. Let $G = (V_1, V_2, V_1 \times V_2, w)$ be a weighted bipartite graph such that $w : V_1 \times V_2 \to R$. An *induced subgraph* G' of G is defined by subsets of V_1 and V_2 along with *all* their edges, namely $G = (V_1', V_2', V_1' \times V_2', w)$ where $V_1' \subseteq V_1$ and $V_2' \subseteq V_2$. Prove that the problem of identifying the heaviest induced subgraph in G is NP-complete.

b. Let G be a bipartite graph where each VERTEX is assigned a weight (rather than edges as in the previous item). The *maximum node biclique* in G is a complete subgraph such that the sum of weights of its vertices is maximum. Devise a polynomial algorithm for solving the maximum node biclique problem.

5 Finding simple paths and trees (30 pts)

In the following, let $G = (V, E)$ be an undirected graph. We are interested in finding a simple path of length k (k vertices) in G. One way to find such a path is to choose a random acyclic orientation of G and look for a directed path of length k under this orientation. In this question we will analyze the performance of this method.

a. A random acyclic orientation can be obtained by choosing a random permutation $\pi : V \to \{1 \dots |V|\}$ and directing an edge $(u, v) \in E$ from u to v if and only if $\pi(u) < \pi(v)$. Show that a directed path (of length k)

in the resulting graph must be simple, and provide a linear-time dynamic programming algorithm to find one.

b. Compute the probability that a path of length k becomes a directed path under a random acyclic orientation. Conclude that one can find a simple path of length k (if one exists) in $O(k! \cdot |E|)$ expected time using the random acyclic orientation procedure. For what values of k is this method faster than color coding?

c. Color coding can be used to find more general structures than paths. Adapt the color coding method to identify a simple tree of size k (k vertices) in G. Analyze the complexity of the algorithm you suggest.