THE TWITTER PROJECT: MY NOTES

GAUTAM SISODIA

1. To do

- Migrate to MySQL, server

2. Graph database

2.1. **Nodes.** The nodes represent twitter users. A node has as data the user id, user name, number of friends and number of followers.

2.2. **Relationships.** There is a relationship from node A to node B if and only if twitter user A follows twitter user B.

3. Collecting data

3.1. **Sqlite.**

4. Hierarchy

4.1. **The paper** [1]**.** Let $G = (V, E)$ be a directed graph. Write $n = |V|$, $m = |E|$. All of the following definitions are from [1].

Definition 4.1. A **ranking** of G is a map $V \to \mathbb{N}$. Denote the set of rankings of G by $R(G)$.

Definition 4.2. The **agony** of a ranking $r \in R(G)$ is

$$A(r) := \sum_{(u,v)\in E} \max\{r(u) - r(v) + 1, 0\}.$$

The hierarchy of r is

$$h(r) := 1 - (1/m)A(r).$$

Definition 4.3. The **agony** of G is

$$A(G) := \min_{r\in R(G)} A(r),$$

and the **hierarchy** of G is

$$h(G) := 1 - (1/m)A(G).$$

By [1], $A(G)$ is at most m, so $0 \le h(G) \le m$. Also, [1] gives an $O(m^2 n)$ algorithm for computing $h(G)$ and a optimal ranking, namely a ranking $r \in R(G)$ such that $h(r) = h(G)$.

Date: September 10, 2014.

4.2. **Issues with** [1]**.**

(1) The whole of the twitter data collected as of Aug 2 2014 constitutes a directed graph with 2 million vertices and 4 million arrows, so $O(m^2 n)$ isn't good enough. Alternate/approximate algorithms are needed, or MapReduce techniques.

(2) It bothers me that an arrow between vertices with the same ranking adds a positive amount, namely 1, to the agony of the ranking. It is perfectly natural for two people of the same social class to associate. On the other hand, there are legitimate concerns to simply removing the 1 from the definition of agony. Namely,

   (a) the trivial ranking that gives all vertices the same rank is optimal. Perhaps the total number of ranks should be a parameter.

   (b) the current defintion rewards finding a better ranking over settling for an ok one

(3) There may be a mistake in the paper. Consider the following graph G: The algorithm in [1] gives ranks 0, 1, 0, 2 to vertices 0, 1, 2, 3 respectively. This rank has agony 4 (2 from $1 \to 2$ plus 2 from $3 \to 1$). But this ranking is not optimal! The ranking 0, 1, 1, 2 has agony 3 (1 from $1 \to 2$ plus 2 from $3 \to 1$).

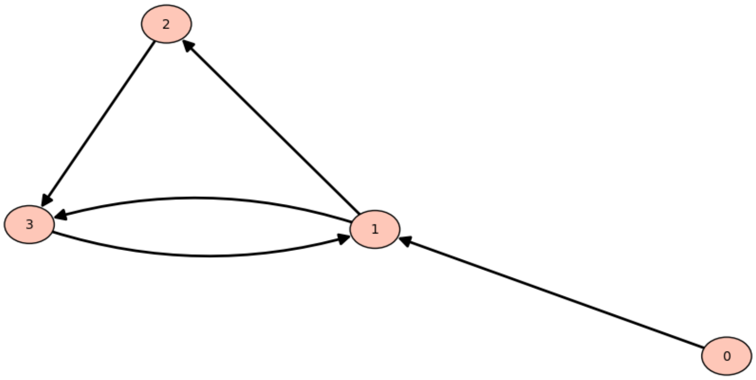

FIGURE 1. Bad graph?

   UPDATE: I'm really convinced that this is actually a mistake in the paper. The mistake is at the end of Algorithm 2, when values are assigned to the integer program variables $x(i, j)$. The problem is that the values assigned may be negative, which is not feasible!

(4) There is another mistake in the paper! Algorithm 2 may not (and usually doesn't) terminate! Consider the graph in figure 2. A maximal eulerian subgraph is $3 \to 0 \to 2 \to 3$. Algorithm 2 will cycle between increasing the ranks of 1, 3 and 0.

4.3. **Hierarchy from a difference matrix.**

Here's another approach to finding a good ranking. Given two vertices u and v in G, compute somehow

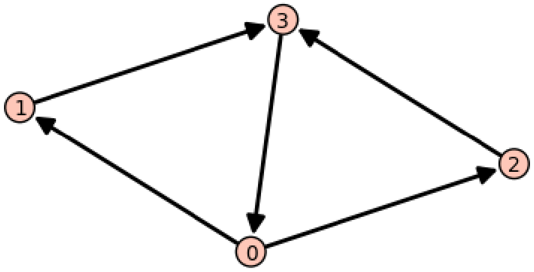

Figure 2. Another bad graph

the best difference $r(v) - r(u)$ (probably by consider the lengths and number of paths between u and v). Form the (non-Euclidean) difference matrix from these ‘distances’. Then use multidimensional scaling techniques to construct the best ranking, which leads to the next subsection.

4.4. **Multidimensional rankings?**

5. A stochastic gradient descent algorithm

References

[1] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, L. Iftode, *Social hierarchy in directed online social networks.*