Dan Sheldon March 14, 2007

Research Summary

My interests lie at the intersection of algorithms, networks, information and knowledge discovery. I am currently working on three projects:

1. Developing link analysis techniques for detecting web spam and inferring content ratings for web pages. Joint work with John Hopcroft.
2. An investigation of network evolution in the collaboration networks of two open-source software development communities, using data from a defect-tracking database. Joint work with Jon Kleinberg.
3. Using graph flow techniques to infer bird migration paths from citizen science data; then building an epidemic model for avian influenza that takes migration of wild birds into account. Joint work with Dexter Kozen.

Web Spam and Content Ratings

Link spamming attempts to boost the search engine ranking of a page or group of pages by carefully arranging links to deceive a link-based reputation algorithm. With John Hopcroft, I am investigating a number of link analysis techniques for finding such arrangements, specifically, those designed to boost PageRank. Our goals are twofold: to advance the state of the art in classifying web spam, and to develop reputation schemes that are robust against collusion.

No single heuristic can correctly classify all spam pages, and the best classifier will almost certainly combine text and link features. Our approach is to develop a number of link-based spam heuristics for use as features in a machine learning algorithm. Then, using sensitivity analysis and feature selection, we can evaluate which features are most predictive, feeding back into our design of spam-resistant reputation schemes. Several of our heuristics derive from the following hypothesis: nodes that collude to boost their PageRank rely on a careful arrangement of links and are more sensitive to small perturbations of the calculation than honest nodes. We consider perturbations both of the graph structure and PageRank parameters. Initial experimental results are promising.

We are also exploring several ideas from the theory of Markov chains. It is well known that the PageRank of a node is the inverse of the expected return time of that node in the PageRank random walk. A major strategy of link spammers is to create short directed cycles that trap the random walk, dramatically increasing the probability of short return times, even though the return time will be very long if the walk escapes the cycle (e.g., by taking a random jump). We can limit the effect of short cycles by conditioning on the return time being at least a certain length. Alternately, a similar quantity from Markov chain theory is the *hitting time* of a node — the expected time for the random walk to reach the node when starting from a random node. Intuitively, short cycles containing the target node have little effect on hitting time. We believe that these measures will be qualitatively similar to PageRank, but more robust to collusion, perhaps provably so.

Finally, we are pursuing a similar idea for inferring the content rating of web pages. We compute the probability that a random walker starting at a given page hits a labeled set of pornographic pages in a short number of steps. This estimates the "threat level" of the starting page.

Evolution of Collaboration Networks

Recent years have shown considerable interest in large real-world networks, and there has been a productive exchange of ideas among those attempting to measure, model, interpret, or develop algorithms for networks. Most measurement efforts have focused on network snapshots covering a single period of time. In contrast, with Jon Kleinberg, I consider the growth and evolution of two real-world networks of collaboration in open-source software development communities.

Defect tracking tools like bugzilla (`http://www.bugzilla.org`) are the locus of interaction for many software teams, tracking not only defects but feature requests, enhancements and product planning. In addition to records of interaction, there is extensive information describing the nature of interaction, including comments by the participants describing their actions.

This data allows us to address many novel questions. Though most analyses consider only pairwise relationships, our data reveals richer knowledge, for example, the paths followed by bugs as they are delegated and possibly re-delegated from person to person. In this sense we can determine which network paths are *used*, not only which paths exist based on pairwise relationships.

Similarly, the defect data yields a natural measure of importance on individuals, based on the issues they raise, delegate, and resolve. Since both software products started as proprietary software, the networks start as dense cores of the original paid developers, and incorporate community members over time. However, few community members seem to make it into the core, and we are studying the mechanisms that allow them more than superficial entry. This represents another significant departure from previous analyses where global network properties have been relatively constant over time but individual links come and go. Our networks are truly evolving.

Mining Citizen Science Data: Inferring Bird Migration using Graph Flows

An interdisciplinary team of researchers at Cornell has proposed an ambitious project to model the potential spread of avian influenza through wild birds. A key ingredient in the model is quantitative information about bird migration, but surprisingly little is available in the scientific literature.

I am working with Dexter Kozen to mine abundance estimates and migration paths from a huge repository of bird observations hosted by the Cornell Lab of Ornithology. Most observations are collected through *citizen science* — individual birdwatchers submit observations through a web application called eBird (`http://www.ebird.com`). The data is organic and messy; statistical methods for the analysis of scientific surveys do not apply. Instead, machine learning and inference methods are more appropriate.

Our work proceeds in several steps. First we develop abundance estimates for all species at all locations over time, leveraging existing statistical and machine learning techniques which we are also extending. Our main contribution is the inference of migration paths. Using abundance estimates, we formulate a flow problem to infer travel between successive time periods. Each location has supply equal to the number of birds at time t and demand equal to the number of birds at time $t + 1$, and uncapacitated edges link all pairs of locations. Then any feasible flow describes a set of flight paths that fit the observations. Given prior estimates of the probability that a bird travels from x to y, we construct a cost function on edges such that the minimum cost flow yields the maximum likelihood flow under the prior. From the flow solution, we extract "typical" migration paths for comparison with those known to ornithologists. Ultimately, we will use abundance and migration estimates as inputs into a stochastic epidemic model.