Detecting Social Ties and Copying Events from Affiliation Data

Lisa Friedland
Department of Computer Science
University of Massachusetts Amherst
140 Governors Dr.
Amherst, MA 01003
lfriedl@cs.umass.edu

Overview. The goal of my work is to detect implicit social ties or closely-linked entities within a data set. In data consisting of people (or other entities) and their affiliations or discrete attributes, we identify unusually similar pairs of people, and we pose the question: Can their similarity be explained by chance, or it is due to a direct ("copying") relationship between the people? The thesis will explore how to assess this question, and in particular how one's judgments and confidence depend not only on the two people in question but also on properties of the entire data set. I will provide a framework for solving this problem and experiment with it across multiple synthetic and real-world data sets. My approach requires a model of the copying relationship, a model of independent people, and a method for distinguishing between them. I will focus on two aspects of the problem: (1) choosing background models to fit arbitrary, correlated affiliation data, and (2) understanding how the ability to detect copies is affected by factors like data sparsity and the numbers of people and affiliations, independent of the fit of the models.

Problem Description. This task first arose within a project to predict fraud in the securities industry, using data provided by the regulatory authority FINRA (Friedland and Jensen 2007). The affiliation data consisted of people and their complete employment histories: for each employee, we knew every branch office where they had worked, along with the start and end dates. The phenomenon we wanted to detect was described informally as workplace "tribes:" small groups of people who "moved together" through their careers, following each other from job to job. This behavior was conjectured to be an indicator for fraud. Our approach was to identify pairs of people who had been co-workers at unusually many, or unusual combinations, of jobs; that is, people who would be unlikely to have so much in common if they had chosen their jobs independently.

I generalize this behavior to other data sets by describing it as one of partial "copying:" in domains where most people act independently of each other, certain people who are closely tied may display coordinated behavior or mimic each other's habits. If we can detect these occurrences, it is a means of identifying *specific* social ties within data that do not explicitly record them. More broadly, although we focus here on people holding the same jobs at the same time, there are countless other rare, specific patterns of correlated activity one could aim to detect; many such links could be useful to analysts preventing fraud or uncovering covert networks.

In another domain, this task could help biologists reconstruct animal families from co-occurrences in sightings of herds (Cairns and Schwager 1987). If the data consist not of people and affiliations, but instead of documents and words, then the copying problem is like plagiarism detection among documents; if the data are entities and attributes, then the copying problem resembles a special case of record de-duplication.

For most such existing problems, the similarity function between two entities is chosen on an *ad hoc* basis and justified based on its performance for a task. My approach differs by being hypothesis-based: given models of the copying behavior we seek and of normal behavior, the similarity judgment derives from their likelihood ratio. (Note that the model will describe a "source" and a "copier," but in practice they may be indistinguishable; I mean to describe any situation where there are two individuals but only one decision-making process.)

To illustrate the issues involved, imagine a bookstore clerk observing customers' purchases over the course of a day. She notices that Jeff and Judy come in separately but buy five of the same books as each other; perhaps they are friends or are in a book club together. How likely are they to know each other? Intuitively, this probably depends on:

1. How many customers visited the store that day? (If many, then overlaps are likely to occur by chance.)
2. How many of the store's customers know each other? (If many—e.g., if the store is in a small town—then their tie is more likely.)
3. How many books does the store offer? (If only five, then it is less surprising to see this overlap.)
4. How many books do customers tend to buy? How many other books did Jeff and Judy buy? (If thousands, then it is less surprising to see the overlap.)
5. Which particular books did they share?
   (a) Example: five Harry Potter books. (Popular and often sold as a series, so the overlap could occur by chance.)
   (b) Example: five best sellers. (All popular, so the overlap might still occur by chance.)

(c) Example: five specialized and unrelated books. (None popular, none related; unlikely to happen by chance.)

(d) Example: five obscure yet related titles. (A strange coincidence, but the customers could simply have the same niche interest.)

Previous Work. In the securities industry project, we framed the task as anomaly detection. We aimed to identify pairs of people whose list of shared jobs was unusually low-probability; such overlap, we reasoned, could only have arisen if the people intentionally coordinated their jobs. In order to compute probabilities, we needed to learn a background model of "normal" movement through careers. We had success using a modification of a Markov process to describe these typical trajectories.

The key contributions of this stage were to:

- Formulate the loosely-described "tribes" concept as one of detecting unusually similar people.
- Describe how this task of identifying latent social ties could be useful in other domains.
- Employ a Markov process to flexibly describe the behavior of most people in the data set without specialized domain knowledge.
- Develop several techniques for indirectly evaluating our algorithm's success when the true labels are unavailable.
- Establish that the "tribes" phenomenon is associated with high-risk individuals in the securities industry.

What propel this work forward are two aspects that are not well understood from the initial phase. First, the Markov model cannot exploit all the information about job timings. I plan to proceed, perhaps surprisingly, by entirely ignoring the temporal features. Solving the problem in this more general case will allow me to deliberately re-incorporate timing and other information later, and will meanwhile enable me to address analogs of the problem in non-temporal domains. Second, like Harry Potter books, some sets of jobs are frequently seen in combination, so the Markov model incorporates correlations among jobs. However, a much simpler model, which simply counts the number of jobs a given pair of people shares, also performed competitively. I would like to understand whether that relative performance reflects shortcomings of our model, or whether, on the other hand, it is an inevitable result of the sparsity and high dimensionality of the data.

Proposed Plan. Three elements guide my proposed approach. First, I wish to explicitly define not only a model of normal behavior, but also a model of copying. Then we can distinguish them using Bayesian inference. Writing down both these models lets us reason more carefully about the effects of our assumptions and how these vary in different instantiations of the problem, as well as check these assumptions in data.

Second, is the insight that we can write down and fully solve a simplified version of the copying problem. For example, suppose that the "normal" distribution is a scatter of points along a line, a univariate Gaussian, and the copying model uniformly picks a point and creates a near-duplicate a small distance $< \epsilon$ away. We can analyze our ability to detect these copies and understand how it varies with the numbers of data points and copiers, the locations of the points that are copied, and the uniformity of the "normal" distribution. Reasoning about this low-dimensional problem reinforces the intuition that a key issue is density estimation. Judgments about copying depend on the distance between the pair of points, but also on the number of other points in the region where they lie. In high dimensions, the "curse of dimensionality" will make data more sparse, and we need to understand how this affects the task.

The third element is a class of models that seem suited to density estimation for this high-dimensional, correlated affiliation data: latent Dirichlet allocation, or one of its descendants, hierarchical Dirichlet processes (Blei, Ng, and Jordan 2003; Teh and Jordan 2010). I seek models that will capture arbitrary correlations among jobs, books, or other objects. These processes have been used extensively for topic models, that is, to describe how documents are linked to words via latent groups called "topics." I have much to do in terms of choosing and developing the right models for the data sets, but I believe these are worth investigating and that their use would be novel for the goal of detecting copying (whether of people's affiliations, of documents, or of database records).

Following the development of theoretical tools, much of my research will be empirical. I plan to work with several additional real-world data sets (not yet finalized), detecting copiers such as: people who watch movies together as visible from their Netflix ratings, people who send each other academic articles as visible in papers saved to their libraries on CiteULike, and documents that have been plagiarized. In addition, I will synthesize data from known models and vary the dimensionality properties of all the data sets by selectively truncating them.

Evaluation of copy-detection in real data is a challenge since labeled pairs do not exist. Two methods are available: first, we can artificially insert partial copies into the data and test our success at retrieving them; second, we can exploit ignored features of the data such as timing and word order to validate the proposed pairs. To evaluate background models, I will examine both how well each model fits the data (via likelihood on held-out entities) and how accurately the copies are detected when using that model.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Cairns, S. J., and Schwager, S. J. 1987. A comparison of association indices. *Animal Behaviour* 35(5):1454–1469.

Friedland, L., and Jensen, D. 2007. Finding tribes: identifying close-knit individuals from employment patterns. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 290–299. New York, NY, USA: ACM.

Teh, Y. W., and Jordan, M. 2010. *Hierarchical Bayesian Nonparametric Models with Applications*. Cambridge, UK: Cambridge University Press.