Finding Tribes:
Identifying Close-Knit Individuals from Employment Patterns

Lisa Friedland and David Jensen

Department of Computer Science
University of Massachusetts, Amherst

Overview

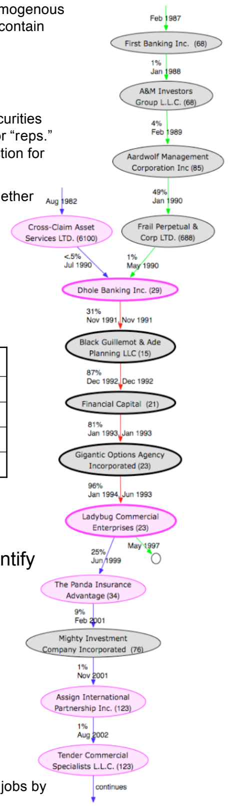
In the securities industry, fraud can be perpetuated by ***tribes*** of employees colluding at multiple jobs. We present a family of algorithms that uses employment histories to detect such tribes: ***small groups of individuals sharing unusual sequences of affiliations***. We treat this as an anomaly detection task and develop models describing typical vs. atypical job transitions within the industry. The resulting tribes tend to be homogenous with respect to risk scores and geographically mobile, and they contain individuals at high risk for fraud.

Motivation

National Association of Securities Dealers (NASD) oversees securities firms in the United States and their registered representatives, or "reps." Responsible for preventing, identifying, and taking regulatory action for cases of fraud.

Hypothesis: Colluding groups of reps, or *tribes*, often move together through multiple places of employment to commit fraud.

Task: Find such groups.



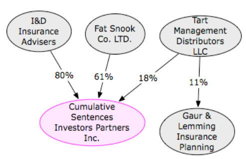
Data Characteristics

1. Employments are not sequential.

Rep	Branch ID	Start Date	End Date
John A. Doe	107	Jan 1985	Oct 1987
John A. Doe	291	Jan 1985	Nov 2000
John A. Doe	382	Mar 1988	Dec 1988
John A. Doe	107	Dec 2003	present

- Employment stints overlap or leave gaps.
- Reps may take different paths between same jobs.

2. Want to factor out background patterns to identify reps that stay together intentionally.



- Mass movements are common. Reps may share multiple jobs by chance, due to patterns of transitions in the industry, e.g.:
  - Branches open, close, merge, or are acquired.
  - Typical career paths within different cities.

3. Large NASD data set: 4.8 million employment records, 2.5 million reps, 560,000 branch offices.

Basic Method

1. Find all pairs of reps that have ever worked together.
2. For each pair, examine list of jobs they have shared. Decide if job sequence is anomalous [see below].
3. Each set of reps connected by anomalous links ⇒ a tribe.

Scoring/Ranking Functions

Given a sequence of jobs that two reps have shared. Under the null hypothesis of reps moving independently, is it likely to arise by chance?

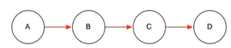
Simple

- JOBS: Count the number of jobs shared. High ↔ unlikely.
- YEARS: Add up time overlapping at each jobs. High ↔ unlikely.

Probabilistic

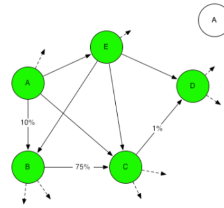
- PROB: Modification of a Markov chain. Low likelihood job sequence ↔ unlikely for two reps to share the sequence.

Markov chain approach



Sequence to score

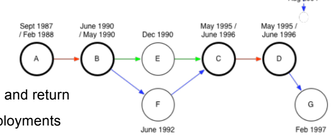
$$P(A \rightarrow B \rightarrow C \rightarrow D) = p_A \cdot t_{AB} \cdot t_{BC} \cdot t_{CD}$$



Transition model

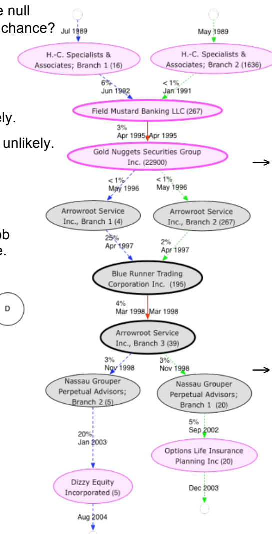
Modifications allow:

- Paths that diverge and return
- Simultaneous employments



Family of probabilistic models

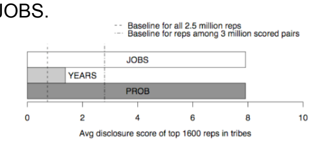
- PROB-TIMEBINS: transition rates differ each year
- PROB-NOTIME: it doesn't matter which job came first



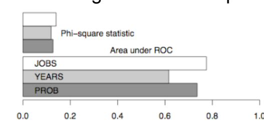
Results

- Scored the ~3 million pairs of reps who had worked at least 3 jobs together.
- Used three scoring functions to rank pairs: JOBS, YEARS, PROB. Figures compare sets of tribes matched to contain 1600 individuals. PROB-TIMEBINS and PROB-NOTIME omitted here, but similar to PROB.

→ Reps in tribes have high risk scores under PROB and JOBS.

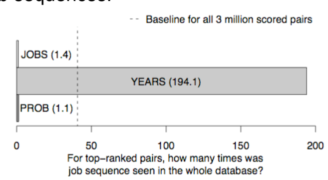


→ Tribes are homogenous with respect to risk scores.

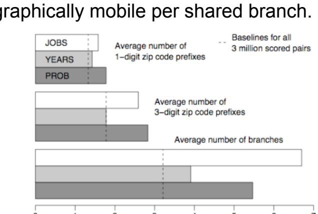


For top tribes: phi-square statistic (normalized chi-square) on presence of risk scores among 2-person tribes; AUC of task predicting risk score as average of tribe-mates'.

→ PROB and JOBS models succeed at identifying rare job sequences.



→ Pairs ranked highly by PROB models are more geographically mobile per shared branch.



For top-ranked pairs: How many cities (zip codes) seen?