Using hubs and authorities to characterize international relations

Michael Friedman and Richard Stoll and Devika Subramanian

Rice University

Outline

- The problem: mining events data
- Our approach: representing events data as interaction digraphs
- Tracking hubs and authorities in interaction digraphs over time
- Validating the hubs and authorities analysis on the Levant event data set
- Summary of results

Events data

Date	Actor	Target	Weis Code	Wies event	Goldstein scale
790415	ARB	ISR	223	(MIL ENGAGEMENT)	-10
790415	EGY	AFD	194	(HALT NEGOTIATION)	-3.8
790415	PALPL	ISR	223	(MIL ENGAGEMENT)	-10
790415	UNK	ISR	223	(MIL ENGAGEMENT)	-10
790415	ISR	EGY	31	(MEET)	1
790415	EGY	ISR	31	(MEET)	1
790415	ISRMIL	PAL	223	(MIL ENGAGEMENT)	-10
790415	PALPL	JOR	223	(MIL ENGAGEMENT)	-10
790415	EGY	AFD	193	(CUT AID)	-5.6
790415	IRQ	EGY	31	(MEET)	1
790415	EGY	IRQ	31	(MEET)	1
790415	ARB	CHR	223	(MIL ENGAGEMENT)	-10
790415	JOR	AUS	32	(VISIT)	1.9
790415	UGA	CHR	32	(VISIT)	1.9
790415	ISRGOV	ISRSET	54	(ASSURE)	2.8

177,336 events from April 1979 to October 2003 in Levant data set (KEDS).

Mining events data

- What can be predicted from events data?
  - Patterns in Israel-Palestine conflict (Schrodt et. al. 2004).
  - Forecasting conflict in the Balkans using HMMs (Schrodt et. al. 2000).
  - Early warning of conflict in S. Lebanon using HMMs (Schrodt et. al. 1997)

The approaches

- Learn statistical models from events data.
  - Example: Sequential data models such as HMMs implicitly build conditional probabilities of observing a particular event type or a conflict level based on history of past events.
- Posit high-level interaction patterns and look for evidence in the events data.
  - Tit-for-tat, olive branch, etc..

Problems with learning approaches

- Overall accuracies are low (~ 50%).
- False positive rates are high (~ 70%).
- False negative rates are high (~ 10%).

Source: **Forecasting Conflict in the Balkans using Hidden Markov Models, Schrodt et. al. 2000.**

Problems with pattern-based approaches

- Need patterns to be defined *apriori*.
- No mechanism to create new patterns directly from the data.
- Most patterns are deterministic not stochastic.

Fundamental problems -- I

- Events data is highly noisy
  - Noise arises from errors made by automated coding programs.
  - Noise arises from "irrelevant" events, events that should not be included in the analysis based on context.

Fundamental problems -- II

- Events data is highly non-stationary.
  - Trends
  - Singularities or discontinuities
  - Regime shifts

Learning methods

- Have to take these fundamental limitations in the data into account.
  - Need to detect singularities.
  - Need to extract quasi-stationary segments in data for learning models.
  - Need to handle noise (not Gaussian!).
  - Need new validation protocols (old train + test splits of data will not work).

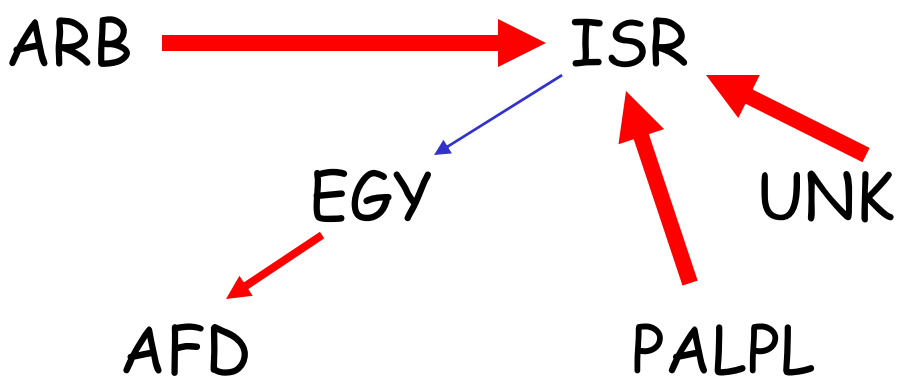
Pattern-based methods

- Need to consider ways of extracting interactions patterns directly from data.
  - What are the patterns on?
    - Generally over actors/targets/events
- This presentation: Using tools from social network analysis to understand temporal variations in key actors/targets in events data.

Key representational idea

- Model interactions between countries in a directed graph.

Date	Actor	Target	Weis Code	Wies event	Goldstein scale
790415	ARB	ISR	223	(MIL ENGAGEMENT)	-10
790415	EGY	AFD	194	(HALT NEGOTIATION)	-3.8
790415	PALPL	ISR	223	(MIL ENGAGEMENT)	-10
790415	UNK	ISR	223	(MIL ENGAGEMENT)	-10
790415	ISR	EGY	31	(MEET)	1



Hubs and authorities

- Given a weighted directed graph $G = (V,E)$, with edges E over a node set V, and weights $A:E \rightarrow Real+$,
  - A hub node is one that has edges directed toward "important" authority nodes.
  - An authority node is one with edges from "important" hub nodes.

Hubs and authorities for events data

- A hub node is an important initiator of events.
- An authority node is an important target of events.
- Hypothesis: Identifying hubs and authorities over a particular temporal chunk of events data tells us who the key actors and targets are!

The HITS algorithm

- Computes hub score (h) and authority score (a) for every node in a weighted digraph.
- Exploits mutually reinforcing relationship between hubs and authorities; and performs an iterative computation to find functions h and a.

The weighted HITS algorithm

- Initialize $h^0(p)$ and $a^0(p)$ to 1 for all nodes in V.
- Iterate until successive h and a estimates are 'close-enough'

$$h^{t+1}(p) = \sum_{p \to q \in E} w(p,q) * a^t(q)$$

$$a^{t+1}(q) = \sum_{p \to q \in E} w(p,q) * h^{t+1}(p)$$

Detail: h and a are length normalized after each iteration

Convergence of weighted HITS

- h and a converge to the eigenvalues of $(w^T w)$ and $(w w^T)$ respectively (Duan et. al, SODA 2004)

Handling non-stationarity

- Picking time granularity of analysis is key to obtaining meaningful results.
  - Divide data too finely (e.g., day by day)-- hubs and authorities are not statistically significant.
  - Divide data too coarsely (e.g., aggregate two months of data) --- hubs and authorities analysis misses key turning points.
  - Ideally, should use another analysis to determine singularities and then decide on data aggregation scheme.
  - In this presentation: we consider events data in chunks spanning two week intervals.

The hypothesis, again

- Hubs and authorities will reveal key actors and targets over a given time period.
- We will use this information to filter irrelevant events (defined as events between countries that are not hubs or authorities)
- We will use simple AR models on hub and authority scores to forecast hub/authority scores.

Analysis protocol

- Take Levant data from April 1979 to October 2003 and chunk into biweeks.
- For each biweek, compute hubs and authority scores using weighted HITS.

Validation using MID data

- Number of biweeks with MIDS in Levant data: 41.
- For each biweek with an MID, check to see if countries participating in MID occur among (top 10) negative hubs and authorities for that biweek.
- Result: For all but 3 biweeks, hubs and authorities analysis predicts MID participants correctly.

Validation continued

- If negative hubs and authorities are included from previous biweek, our accuracy is 41/41!