Dynamic Relationship and Event Discovery

Anish Das Sarma[1], Alpa Jain[1], Cong Yu[2]

[1]Yahoo! Research and [2]Google Research

Roadmap

- Motivation and problem definition
- DROP: Overview of our approach
- DROP algorithm details
  - Relationship detection
  - Event consolidation
- Experiments

Motivation

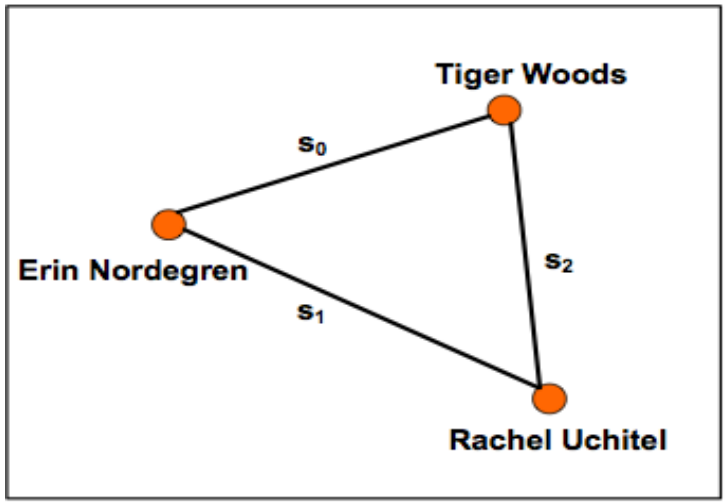
- Examples: Boating accident involving football players, Tiger Woods incident

- Relations in news events are often hard to capture using traditional extraction methods because:
  - Relations may not conform to pre-existing schemas
  - Timeline of involved entities is not constant

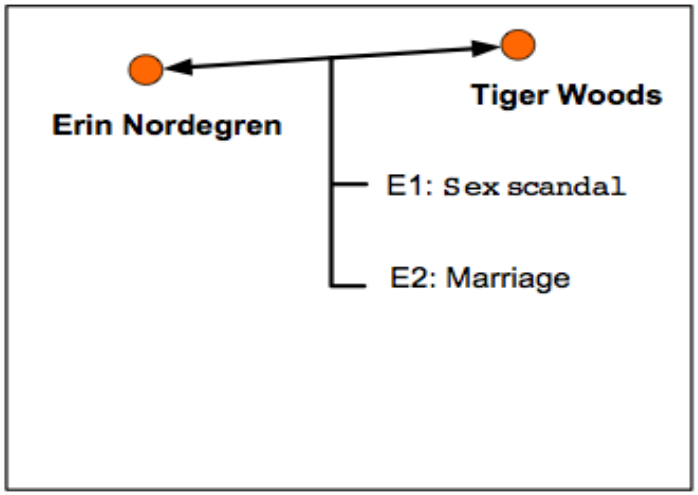
Basic Definitions

- **Dynamic relationship (e1, e2, t):** A pair (e1, e2) of entities *temporally connected* over time interval *t*.

- **Event (N, R, t):** A set of entities N and dynamic relations R ⊆ N x N over a continuous time window t.



Event



Dynamic relations

DROP: Dynamic Relation Problem

Given a set S of entities, a set D of data sources, and a time period T. Find:

1. Every event ***e*** that occurred in a time overlapping T
2. For each event ***e*** compute properties such as:
   - **Entity involvement:** Measure of each entity's involvement in ***e***
   - **Event description:** Keywords describing the event
   - Event popularity/confidence, …

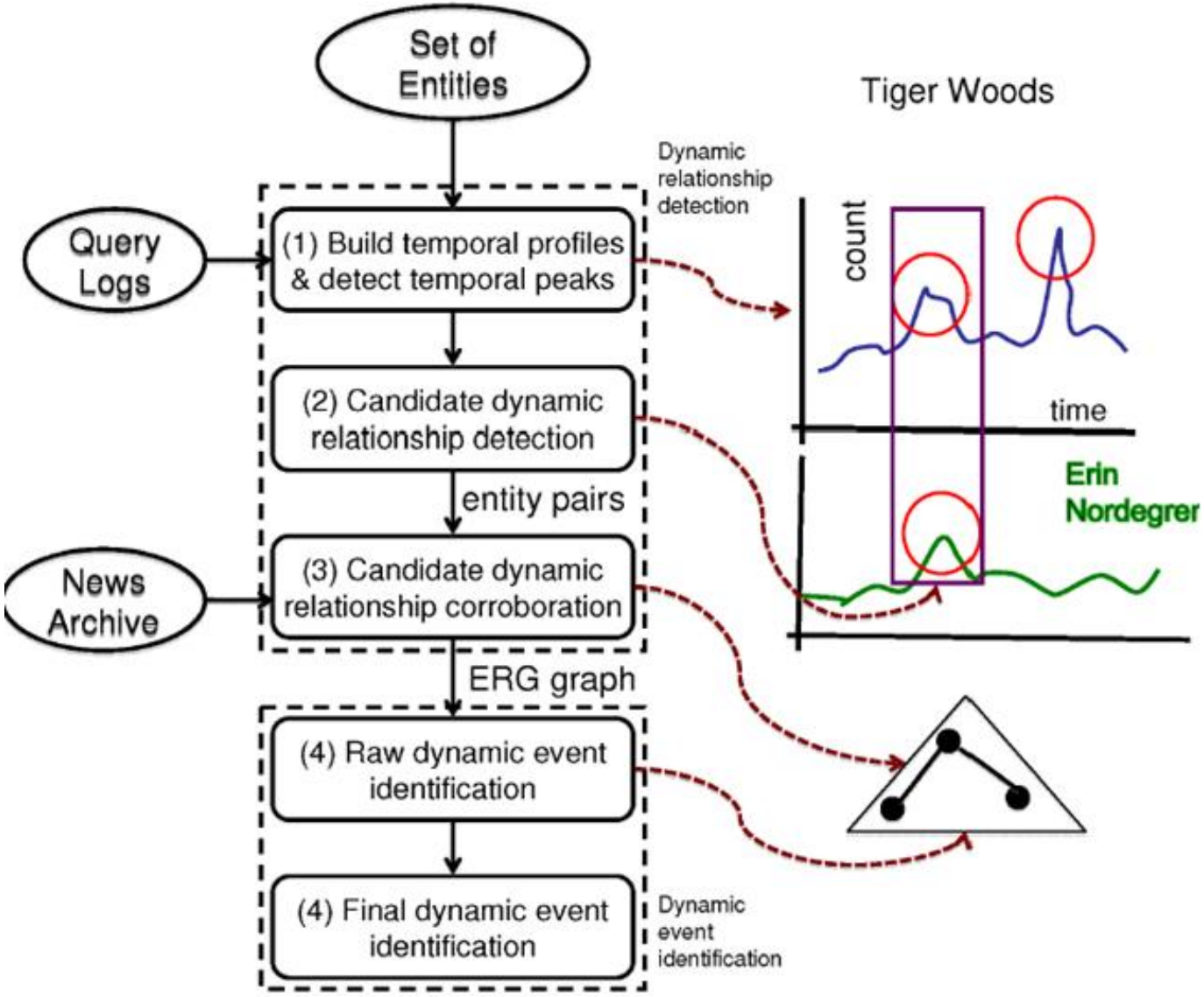
Overview of Our Approach

1. Identify dynamic relation between a pair of entities
   - Build temporal profiles of entities
   - Detect peaking trends in temporal profiles
   - Identify candidate dynamic relationships
   - Corroborate dynamic relation evidence

2. Derive holistic dynamic events from dynamic relations
   - Identify temporally consistent relations
   - *Globally*- and *locally-consistent* clustering

3. Derive event properties (confidence scores, event descriptions, …)

Sample Event Detection



Profile and Peak Detection

- We build temporal raw profile over a time period using frequency of mentions among documents
  - Documents can be twitter feeds, query logs, news archives

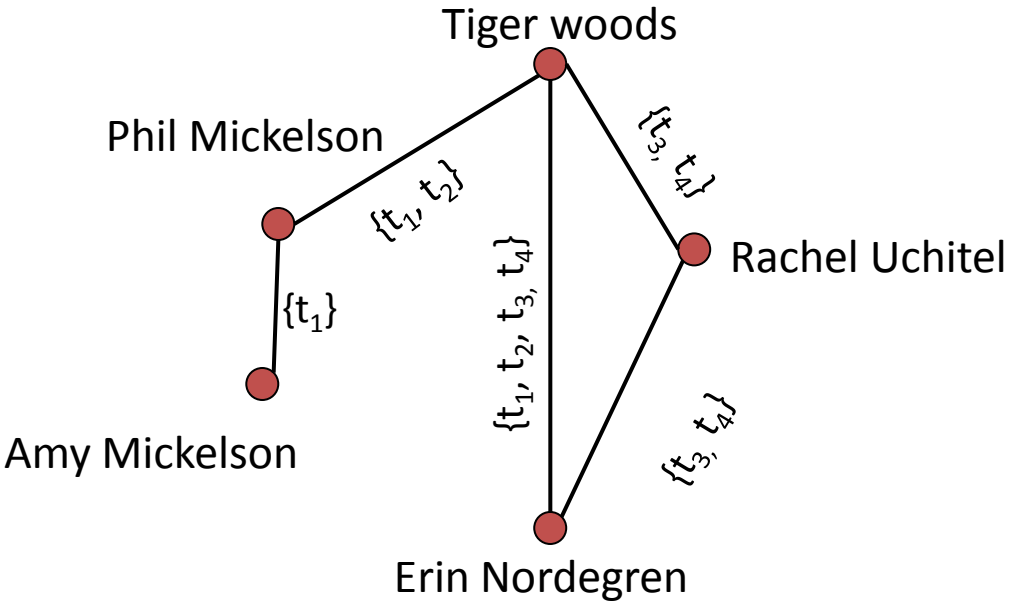
- Two peak detection strategies to identify buzzy entities
  - Rapid Rising
  - Rapid Fall

Candidate Dynamic Relationship

- Identify co-peaking entities
  - An event involves entities with co-peaks or close peaks
  - Entities may co-peak multiple times, each for an event
- **Problem:** Co-peaking may occur coincidentally
- **Solution:** Corroborate evidence from other textual sources
  - Compute association scores PMI between co-peaking entities from different sources

Pair-wise Temporal Graph (PTG)

- Entities as nodes
- Binary relations as edges
- Edge associated with a set of time windows



Events are clusters of connected entities

But,

Maximal clique subject to same week will produce partial events

Connected component with no temporal constraints produces unrealistic large events

Temporally-aware Clustering

- **Input:** Pair-wise temporal graph
- **Output:** Set of clusters, representing events
- Clustering Challenges:
  - Group entities based on dynamic relationships
  - Respect temporal constraints: An event is usually short-lived and cannot span an unbounded time interval

Clustering Goal

Input: Pairwise Temporal Graph: *A pairwise temporal graph (PTG) G = (V,E,W) is an extended EDR graph consisting of a set V of vertices representing entities, a set E of edges representing dynamic relationships, and a function W:E →2^N that associates each edge with a set of time windows; N denotes the finite set of time windows of which the data is present.*

Output: Clustering: Clustering of entities in E such that: (a) entities in E are connected through dynamic relationships, (b) these dynamic relationships are ``temporally-correlated''

Next…

- Clustering
  - Two types of clustering – local and global
  - Intuition + Formal Definitions
  - Example
  - (Algorithms in the paper)
- Event Enrichment
- Experiments

Local and Global Temporal Constraints

- Cluster defined by a connected set of nodes
- Local temporal constraint:
  - Adjacent set of edges are temporally connected
- Global temporal constraint:
  - The entire set of edges are contained in a given global time window

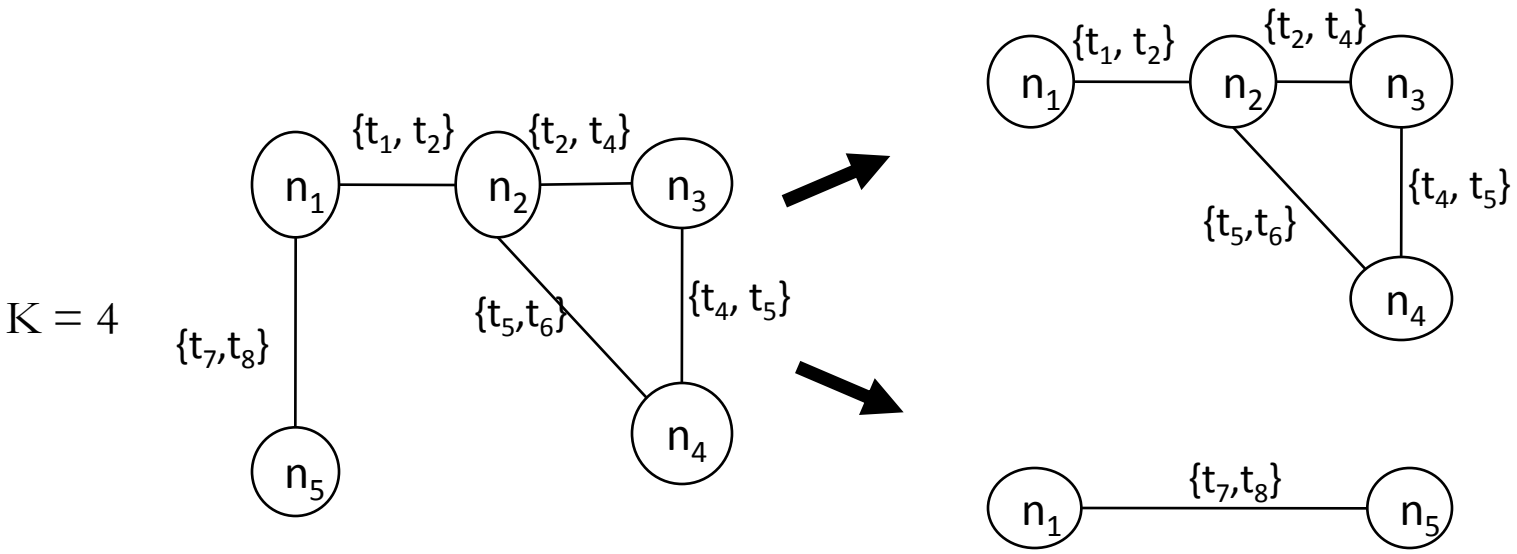
Global Temporal Clustering

GTC Cluster: *Given a PTG G = (V,E,W), and a maximum week-difference parameter K, we say that* C *is global-temporally constrained cluster (GTC cluster) if and only if all the following hold:*

1. ***Connected:*** *there exists a set E_C of edges and a K-week time window W such that:*
    1. *EC connects all nodes in C*
    2. *Every edge in E_C has a time window contained in W.*
2. ***Maximal:*** *Not exists a superset C' of nodes satisfying (1) above.*

Global Temporal Constraints

- Pick a time window K
- Cluster entities such that edges have at least one time interval in K



Local Temporal Clustering

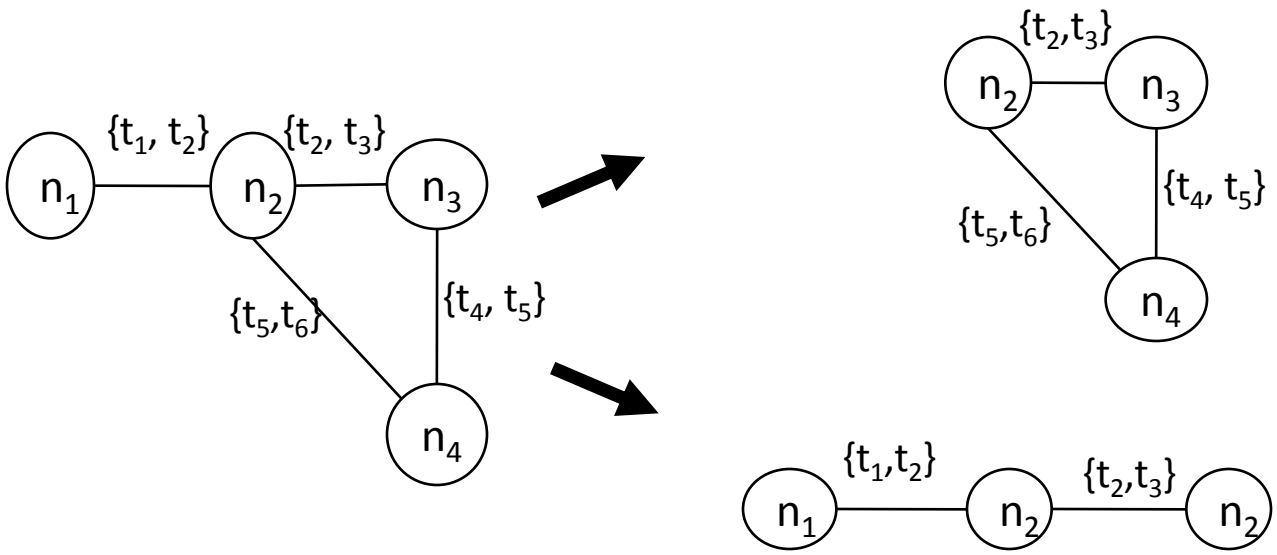
Intuition: C can be connected by a set EC of edges such that every pair of adjacent edges in a spanning tree with edges ET (subset of EC) share a time-window

LTC Cluster: *Given a PTG G = (V,E,W), we say that C is a local-temporally constrained cluster (LTC cluster) if and only if all the following hold:*

1. *The graph GC = (C,EC) is connected, and there exists a spanning tree T=(C,ET) of GC such that:*
2. ***Sharing:*** *For adjacent edges e1, e2 in ET, their time window's are non-disjoint*
3. ***Continuity:*** *There exists a time window $W=[w_{min},w_{max}]$ such that:*
    1. *For any window [w1,w2] in W with |w1-w2|=1, there is an edge e containing W*
    2. *Every edge e in ET has a window in W*
4. ***Maximal:*** *Does not exist a superset C' satisfying the above.*

Local Temporal Constraints

- Events consist of continuous relations
- Cluster entities if edges have at least one overlapping time period



Event Enrichment (summary)

- **Entity involvement:** Strength of the edges incident on the entity
- **Event confidence:** Probability of connectedness of the cluster based on edge weights
- **Event Description:** Most prominent bag of words in documents corresponding to the dynamic relations
- **Event Popularity:** Query logs to determine the number of queries pertaining to the event in the given time window

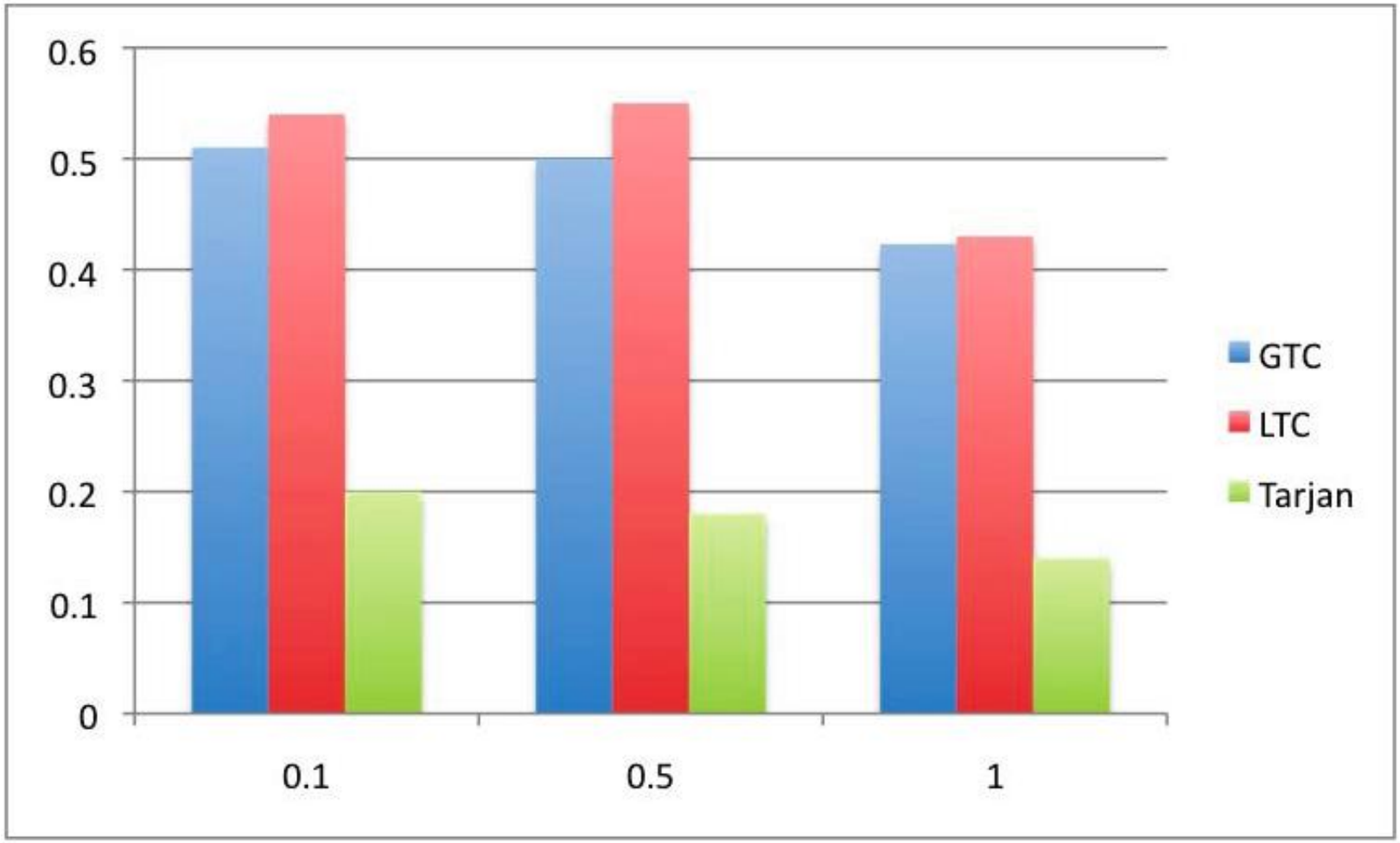
Experiments

- **Dataset:**
  - 100 million anonymized queries; news corpus for 2009, 2010
  - 30K entities using Wikipedia people names
- **Metrics:** Precision, Recall
- **Statistics:**

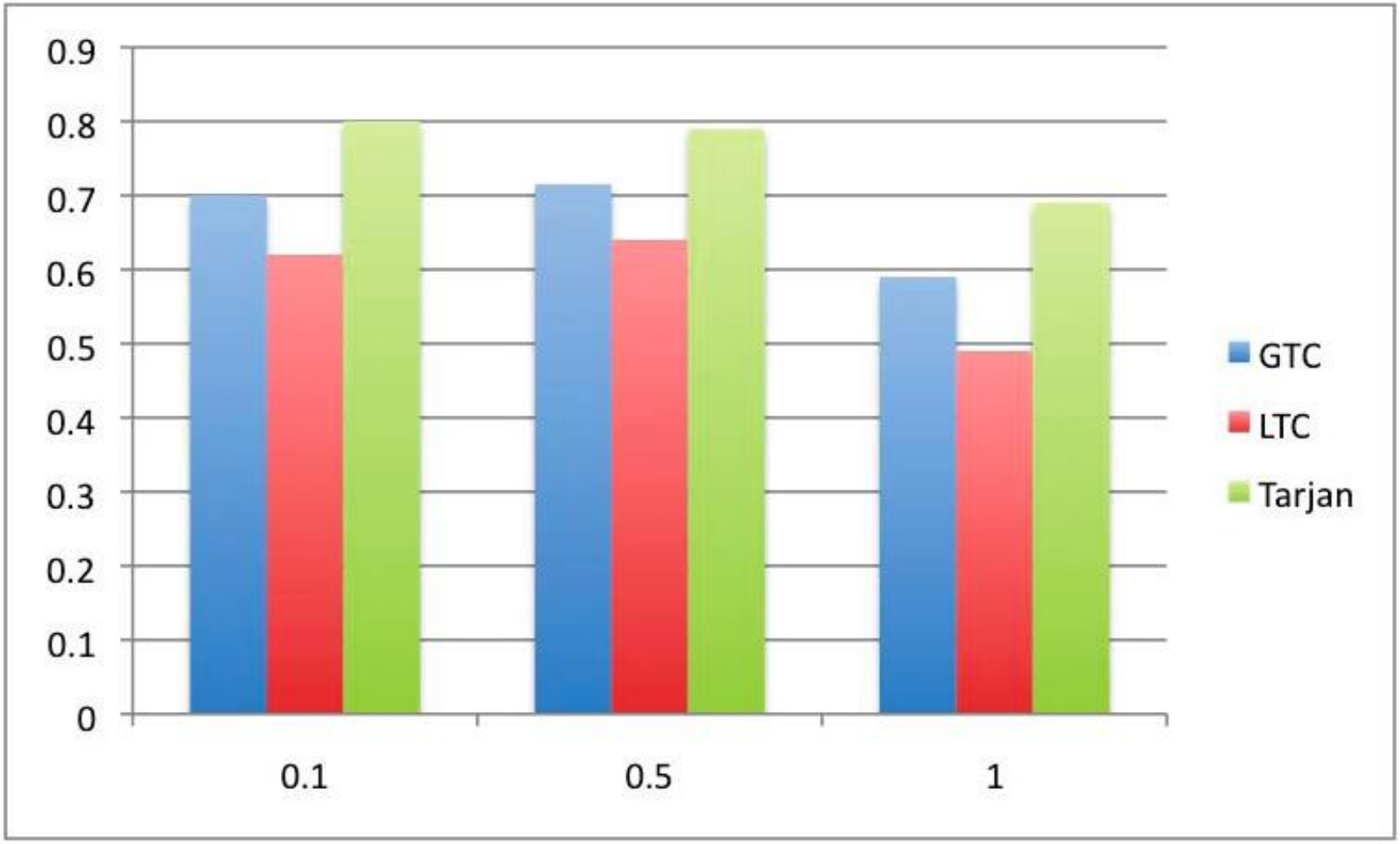
Method	#Events	Average size
GTC	1288	7.22
LTC	1380	4.88
TRJ	120	24.42

Example events detected by DROP

Time	Involved Entities	Event
Week of 02/26/2010	Corey Smith, Marquis Cooper, Nick Schuyler	Boating incident
Week of 01/29/2010	Edward Norton, Gloria, Estefan, Derek Jeter, Alicia Keys	Concert for Haiti
Week of 10/30/2009	Cloris Leachman, Gene Wilder, Mel Brooks, Peter Boyle	Broadway musical tour



Average **recall** of GTC, LTC, Tarjan over entire gold set, varying match threshold



Average **precision** of GTC, LTC, Tarjan over entire gold set, varying match threshold

Related Work

- **Relation, entity co-occurrence extraction:**
  - Open information extraction [Banko et. al.]
  - [Sarkas et. al.]
- **Detecting buzzy entities:**
  - Modeling temporal behavior (e.g., [Kleinberg, KDD 02])
- **Event detection:**
  - [Zhao et. al., AAAI 07] and others
- More in the paper

Summary

- Developed a system (DROP) for dynamic event and relationship detection detection
  - Input set of entities
  - Query logs, news articles
- Basic Approach:
  - Peaking entities
  - Co-peaking -> Candidate relationship -> Corroboration
  - Clustering
  - Event enrichment

Thanks!

Anish Das Sarma
anish.dassarma@gmail.com