Revealing the Hidden Links in Content Networks: An Application to Event Discovery

A. Saravanou*, I. Katakis*, G. Valkanas#, V. Kalogeraki+, D. Gunopulos*
*University of Athens, #Detectica, +Athens University of Economics and Business

Motivation

Social Networks contain valuable information for event detection.

dblp yelp stackoverflow

Events could be disasters, concerts, sports, ...

Example: FIFA 2014 Draw

- 16:48 - FIFA world cup draw in full flow @talksport
- 16:55 - fifa world cup draw starts now! #worldcup
- 17:01 - Easy group for France
- 17:09 - Italy with Uruguay: Group D

FIFA WORLD CUP Brasil (fifa.com)

LiCNo

Our method

LiCNo (Linking Content Nodes)

- Content Network, a dynamic heterogeneous graph (user + content nodes)

Other methods

Graph based

- Interactions between users
- Active subgraphs

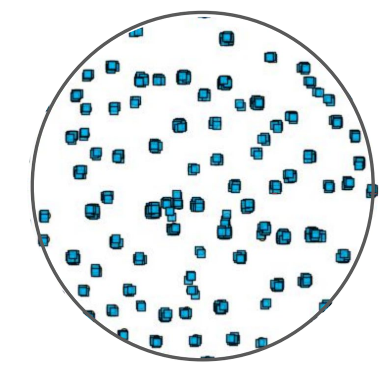
Text based

- Novel context in a stream of text

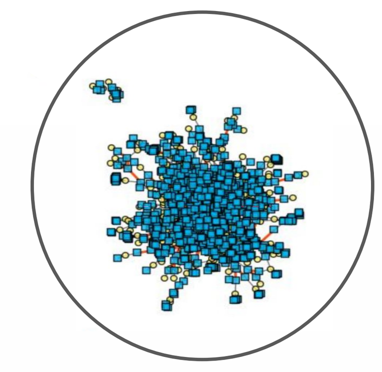
Why Hidden Links?

Hidden links better capture discussions around a topic

Content network *without* hidden links

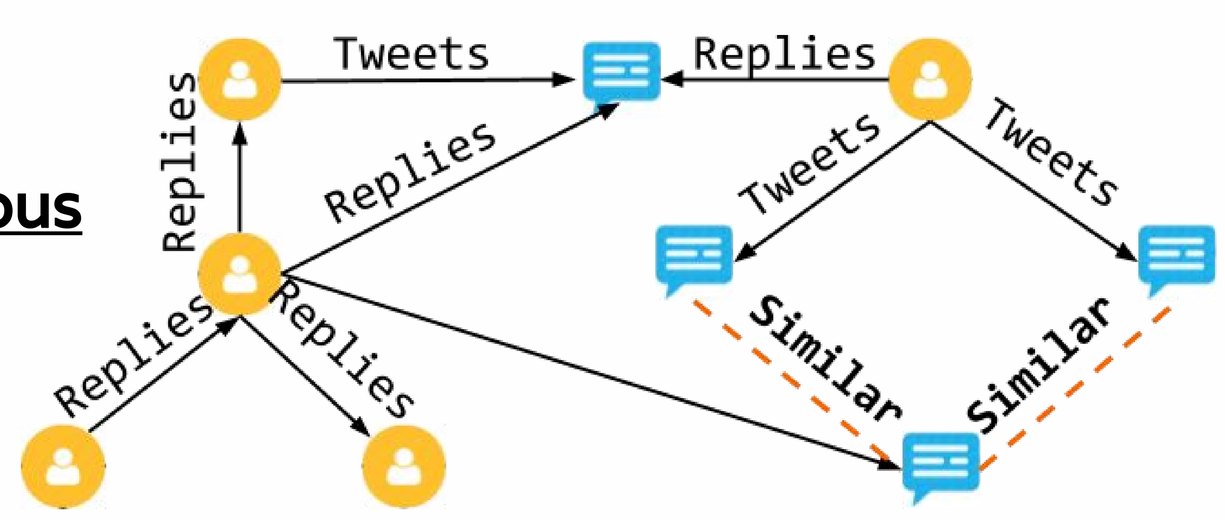


Content network *with* hidden links



Definitions

Heterogeneous graph



i) Snapshot Graph, $G_t = \{ V_t, E_t \}$,
$V_t = \{ V_{(0,t)}, \ldots, V_{(m-1,t)} \}$, where m is the number of different node types, $E_t \subseteq V_t * V_t$

ii) Content Network, $G = \{ G_t \mid t = 1, .., t_{max} \}$,
where G_i is the snapshot graph observed during the *i*-th time window

iii) Event Detection
Given a Content Network, identify a set of events $\mathbf{E} = \{ e_0, \ldots, e_{M-1} \}$, where an event is defined by its description and duration $e_j = \{ d_j, t_{(end,j)} - t_{(start,j)} \}$

Our method

1) Build the *snapshot* graph

- *user*X tweets *text*A
- *user*X replies *user*Y

2) Reveal hidden links

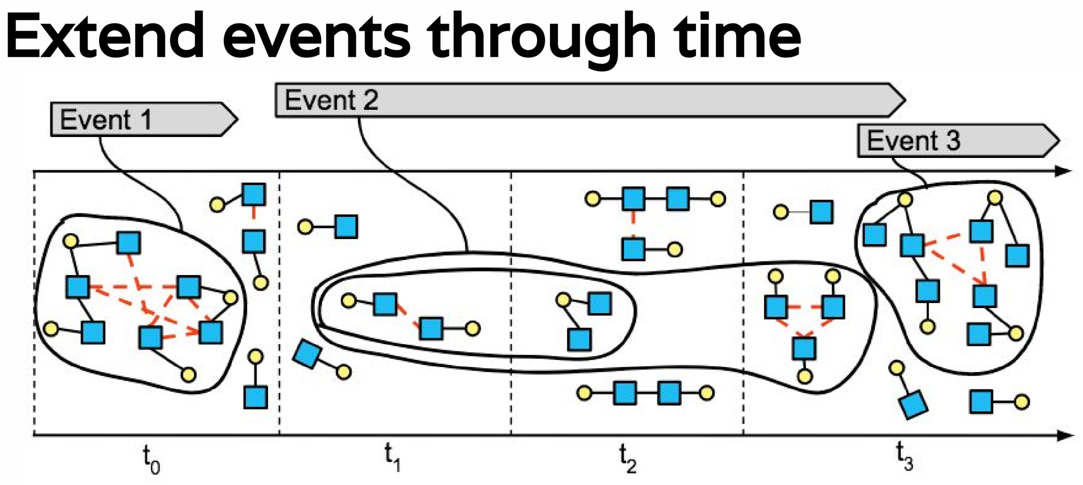
- *text*A is_similar_to *text*B

3) Identify events (very large CCs) **& candidate events** (large CCs)

For all CC_i in G_t:

$$h(CC_i) = \begin{cases} 1, \text{ if } |CC_i| > avg(|CC|) + \theta * std(|CC|) \\ 0, \text{ otherwise} \end{cases}$$

4) Extend events through time

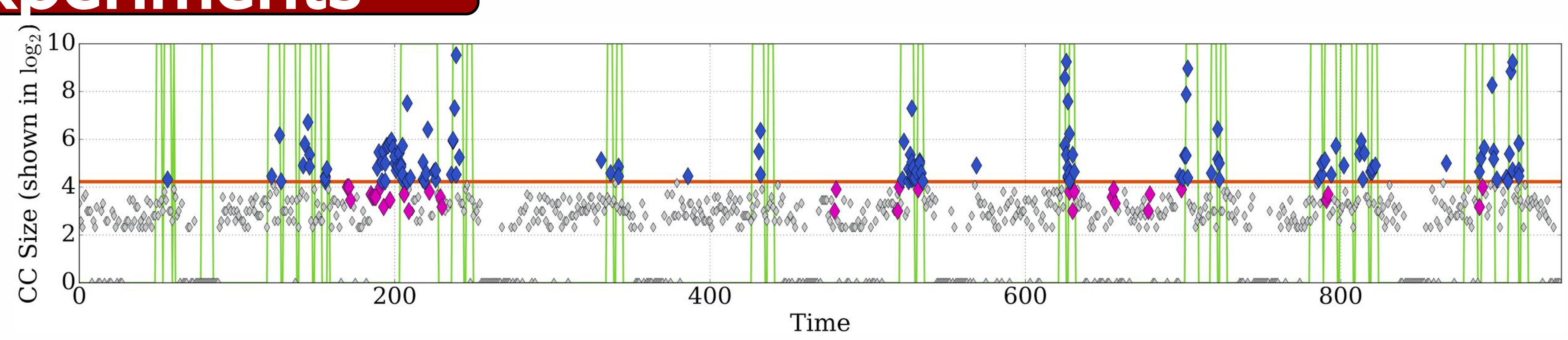


5) Filter

- Spam messages & blacklist incidents

Experiments

1) Dataset: ~ 700K public geotagged **tweets** from London organised into 15-min time windows
Ground truth: Wikipedia & manual annotation



2) Comparison methods:

Baselines:
- *Activity Detector*: unexpected number of tweets
- *Structure Components*: tracks vlCCs on interaction graph
- *Content Components*: tracks vlCCs on content graph

State-of-the-art:
- *SELECT-H*: builds ensembles of anomaly detectors

Our method:
- *LiCNo (tf-idf)*: reveals links using cosine similarity of tf-idf vectors
- *LiCNo (w2v)*: reveals links using cosine similarity of w2v embeddings

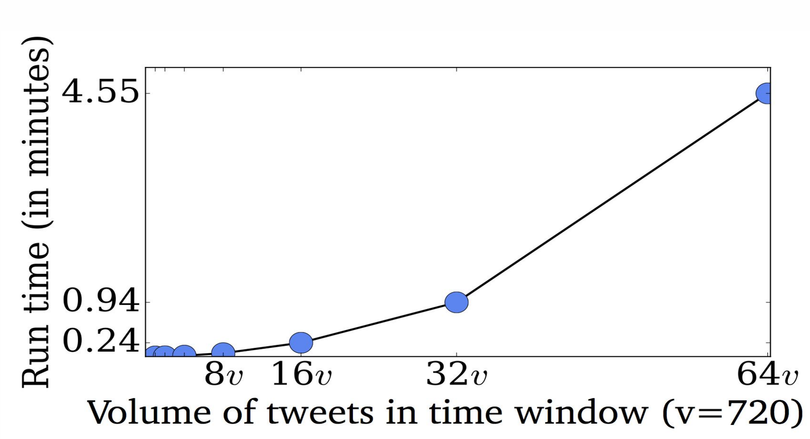
Event Detection

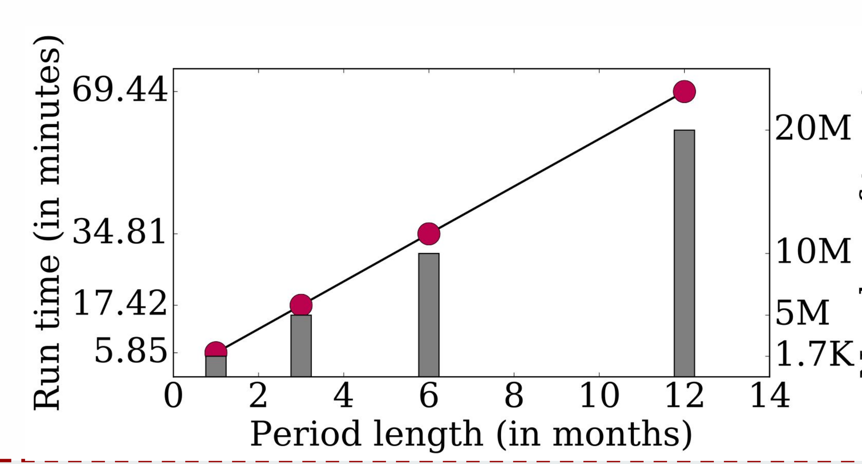
Method	Precision	Recall	F-score
Activity Detector	0.33	0.70	0.45
Structure Components	0.29	$\boxed{0.74}$	0.41
Content Components	0.39	0.49	0.43
LiCNo	$\boxed{0.46}$	0.73	$\boxed{0.57}$

Event Ranking

Method	APrecision	ARecall	AF-Score
LiCNo (tf-idf)	$\boxed{0.65}$	$\boxed{0.69}$	$\boxed{0.67}$
LiCNo (w2v)	0.5	0.61	0.54
SELECT-H	0.3	0.31	0.30

3) Scalability Experiments:
i) Varying volume per time window- left
ii) Varying time period (static volume per time window) - right





detectica

Acknowledgements: This research has been financed by the Horizon2020 688380 VaVeL project (http://www.vavel-project.eu/) , the FP7 ERC IDEAS 308019 NGHCS project (http://pages.cs.aueb.gr/~vana/NGHCS/), and a Google Faculty Research Award.

VaVeL NGHCS Google Faculty Research Awards

CIKM 2017, 6-10 November
Pan Pacific Singapore

contact: antoniasar@di.uoa.gr
www.di.uoa.gr/~antoniasar