FTC 2019

Detecting Spam Tweets in Trending Topics using Graph-Based Approach

Ramesh Paudel, Prajjwal Kandel, William Eberle
Tennessee Tech University

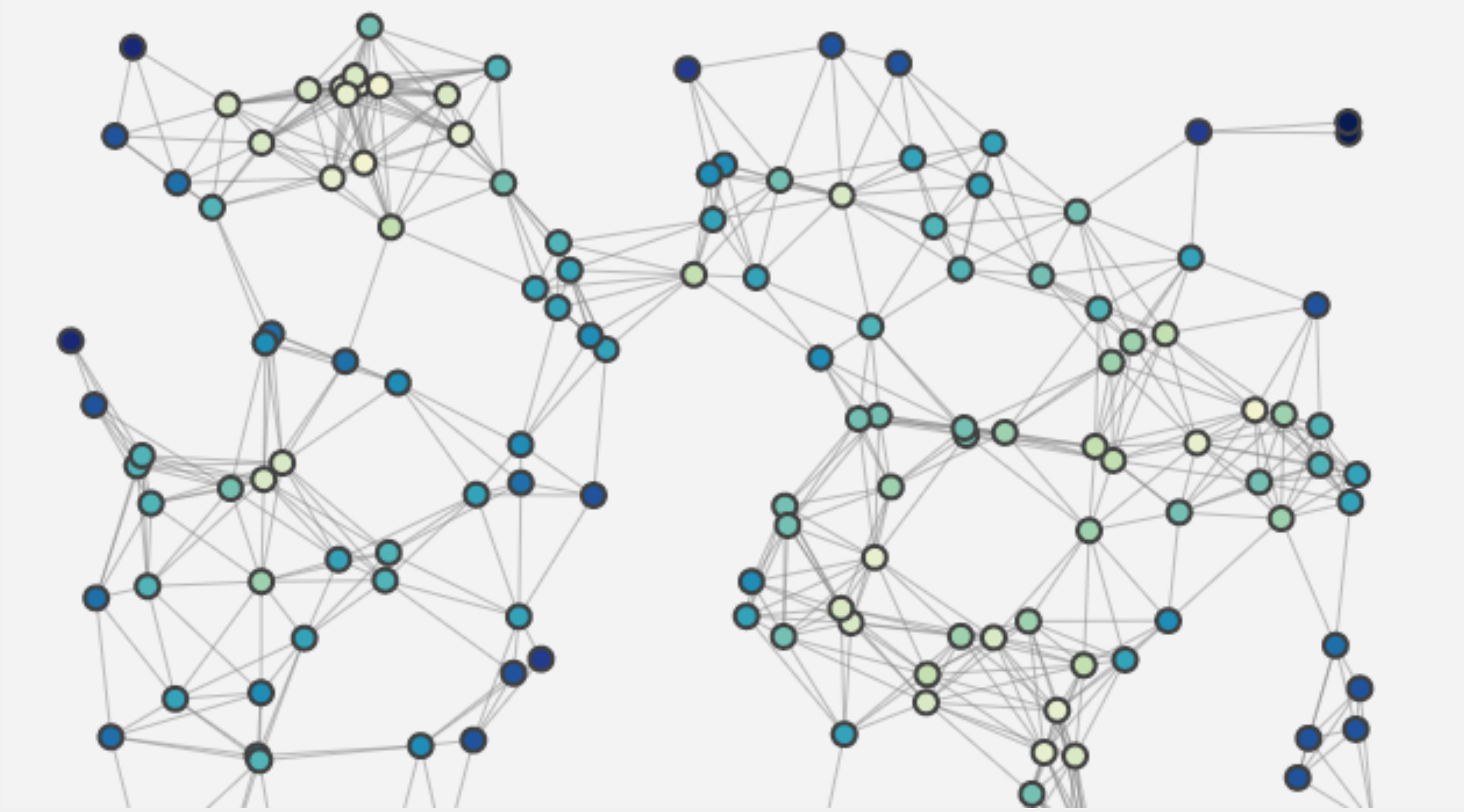
Introduction

- Twitter allows users to post information, updates, opinions, etc., using tweets.
- As the popularity increases around a certain event, more people tweet, thereby making it ”trending”.
- Twitter process is slow, and the trending topics usually last for only a few hours/day at most

- Adversaries use this as an opportunity to propagate off-topic content.
- Spammer may includes a URL, leading the reader to a completely unrelated website.
- Takes advantage of shortened URL.

Our Approach

- An unsupervised graph-based approach
- Extract context of the tweet from heterogeneous sources.
- Context using named entities from the tweets and documents pointed by the tweet.



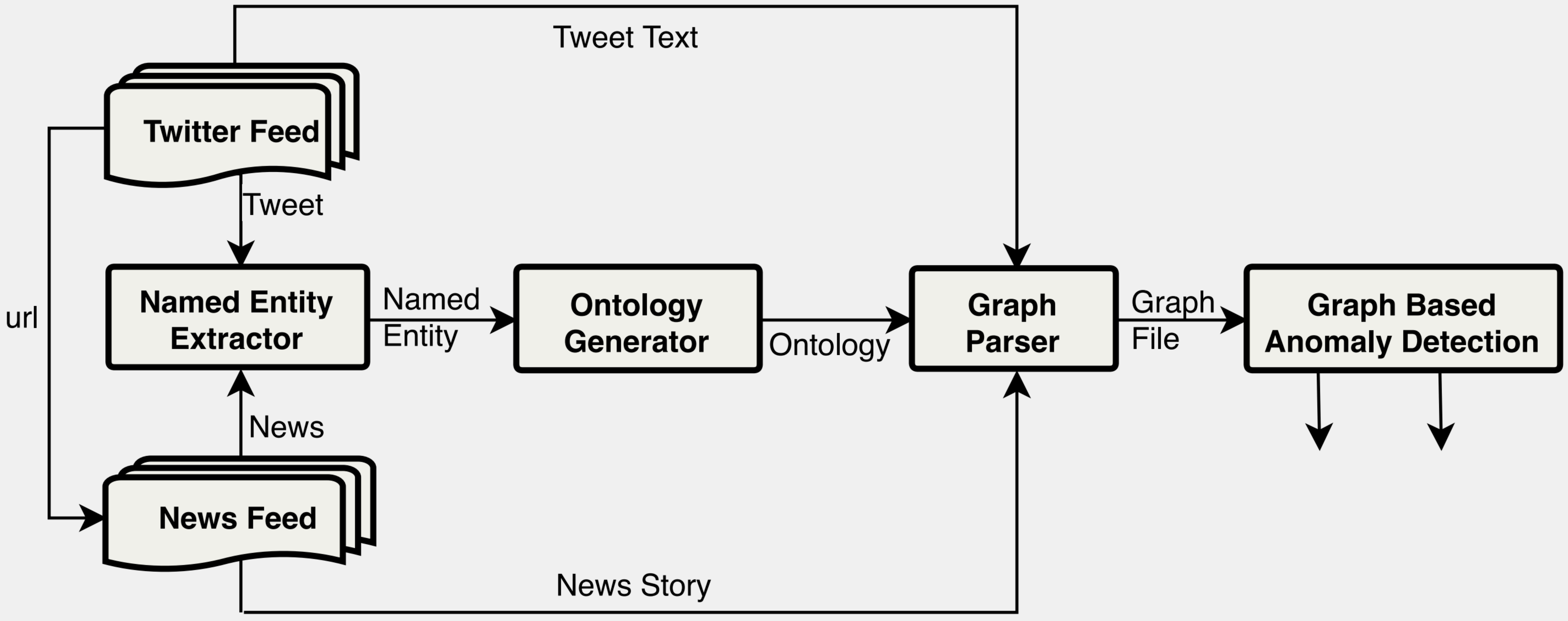
Aim to detect following types of spams/anomalies

- **Keyword/Hashtag Hijacking :-** Using popular keyword/hashtag to promote tweets not related to the topic
- **Bogus Link :-** Posting URLs which have nothing to do with the tweet

Our Approach (cont.)

Four Key Modules:

- **Named Entity Extractor**
- **Ontology Generator**
- **Graph Parser**
- **Graph Based Anomaly Detection**

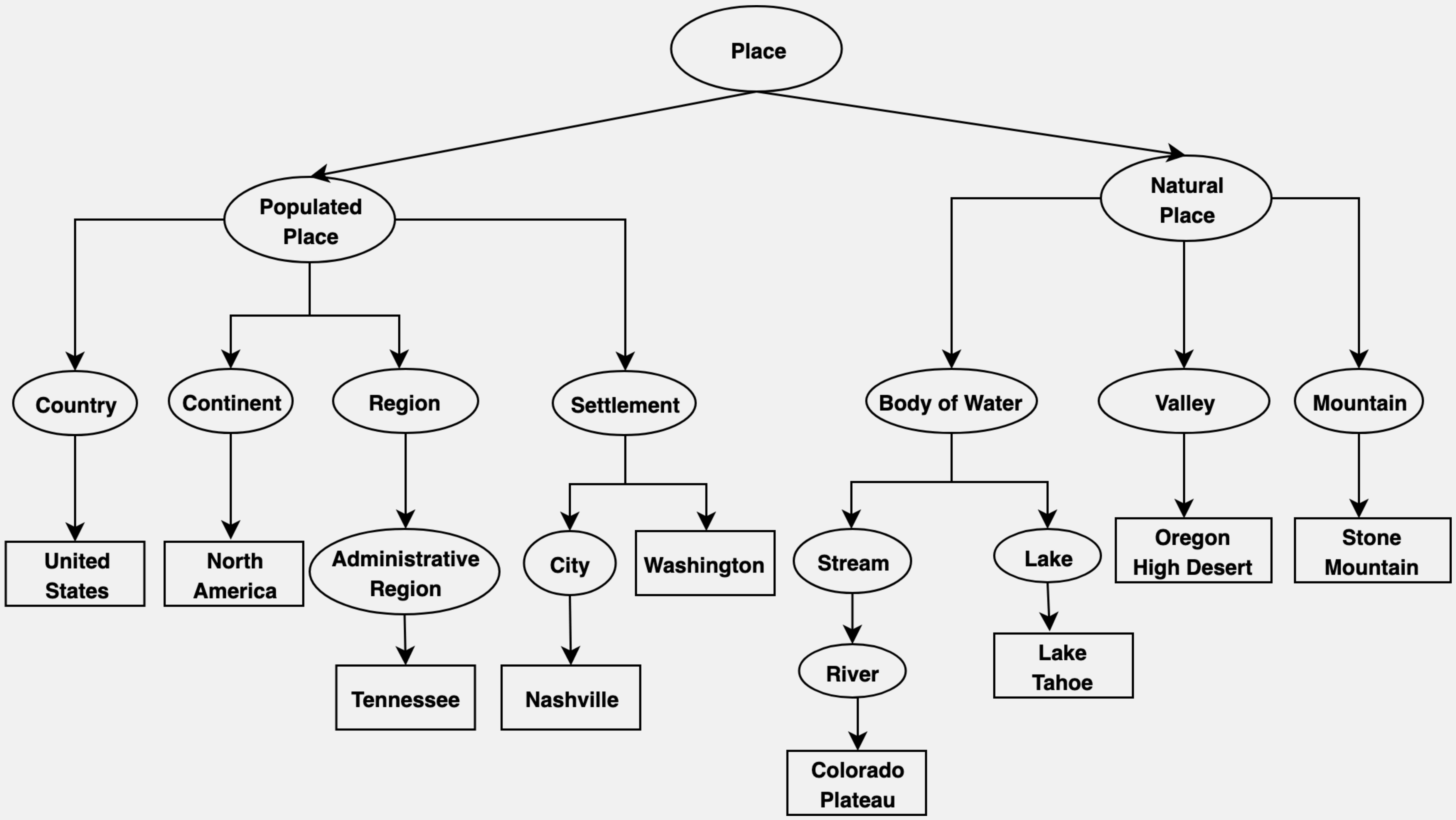


Named Entity Extractor

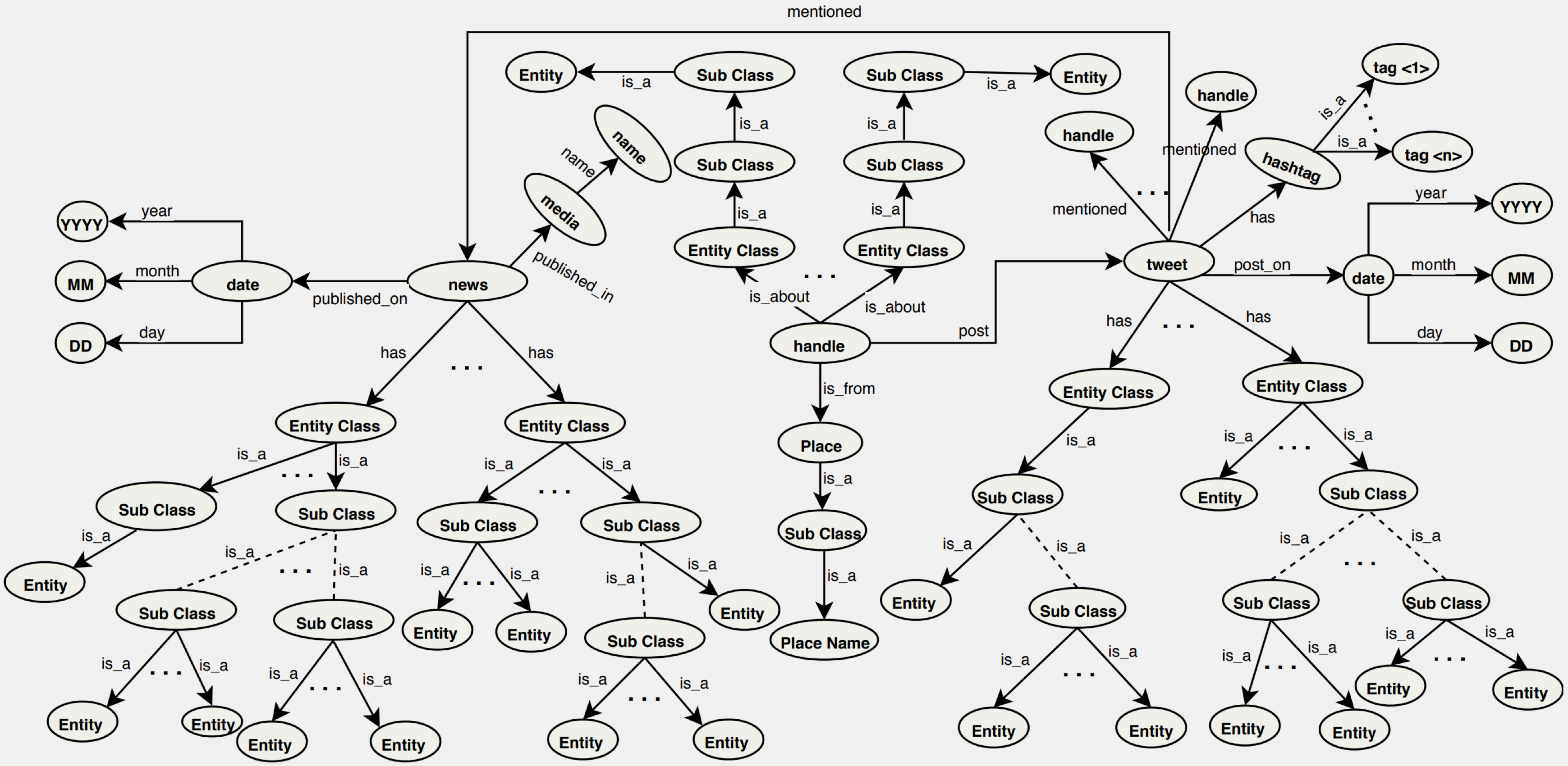
- **Named Entities**: Real-world objects that can be denoted with a proper name like place, person, organization, company, date, etc.

Ontology Generator

- Extracting ontologies provide added knowledge about the named entities
- Used DBpedia API for generating the ontology

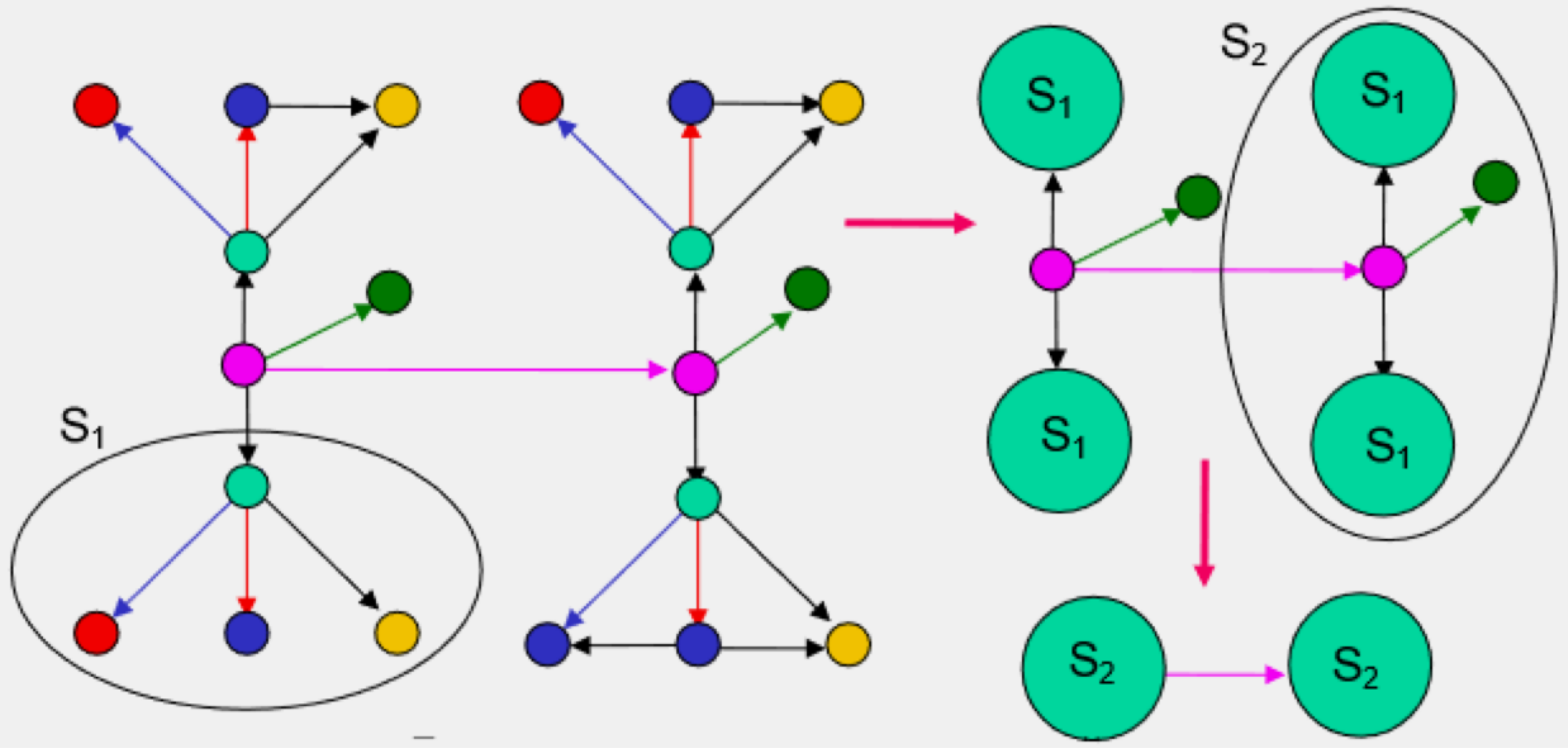


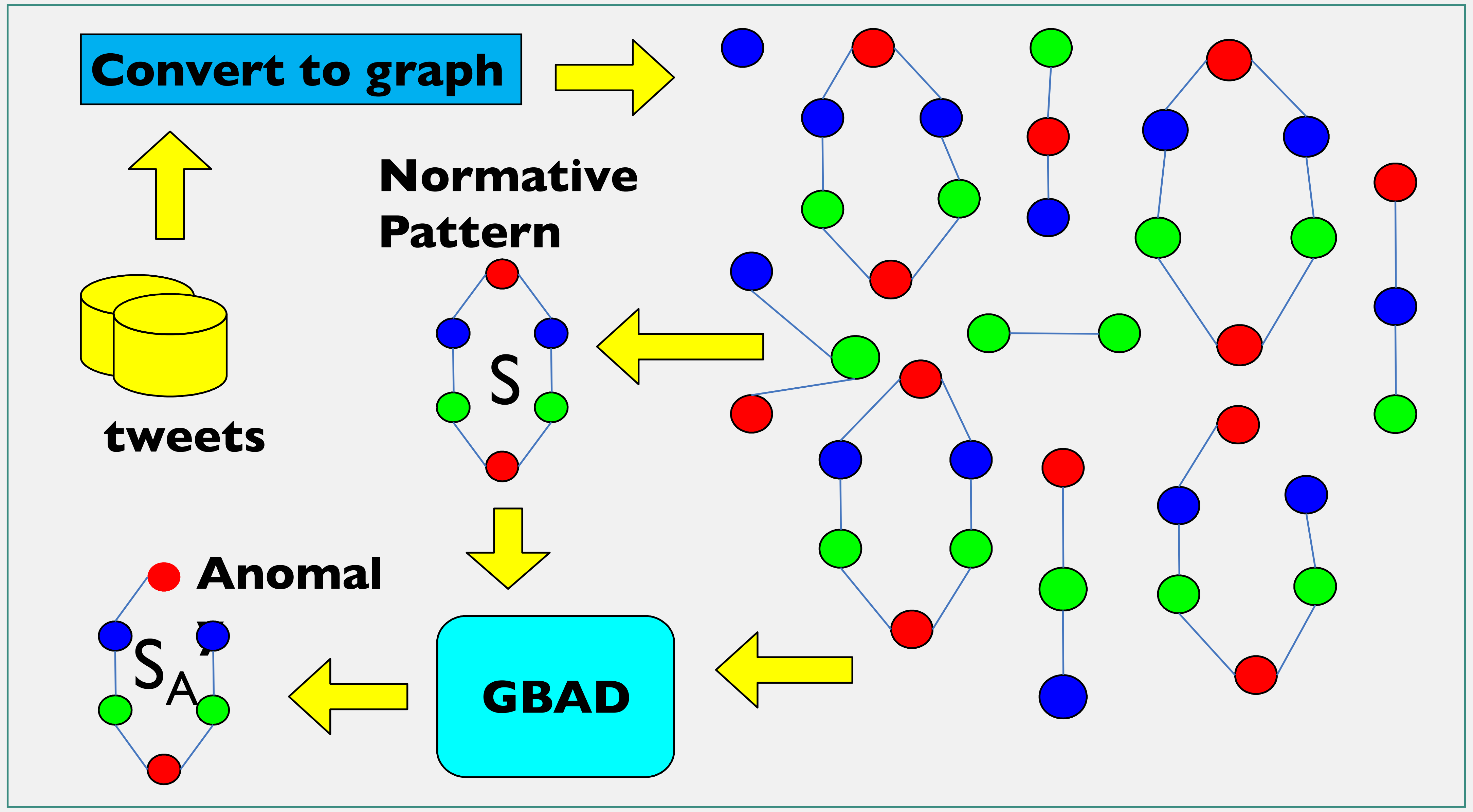
Graph Layout



Graph-Based Anomaly Detection

- Find normative pattern $\boldsymbol{S}$ (highly compressing subgraph using MDL principal)
- Find closely-matching subgraph $\boldsymbol{S_A}$ of $\boldsymbol{S}$
  - Missing nodes/edges (gathered along the way)
  - Additional nodes/edges (search a bit further)
  - Modified labels among structural matches
- $\boldsymbol{P_r(SA)} = \dfrac{\boldsymbol{\#\, particular\, SA}}{\boldsymbol{\#\, all\, SA's}}$
- $\boldsymbol{Anom.score} = \boldsymbol{Pr(SA) * D(SA, S)}$
- GBAD (www.gbad.info)





Dataset

- Collected data using Twitter's standard search API.
- Collected tweets related to two trending topics during the summer of 2018
  - "**FIFA World Cup**"
  - "**NATO Summit**"
- Crawled news from the URLs in tweets

Trending Topic	Total tweet/news	Anomalies		
		Keyboard Hijcking	Bogus Link	Total
World cup	1,463	2	20	22
NATO Summit	1,716	0	11	11

Dataset	Number of Vertices	Number of edges
FIFA World Cup	88,887	87,424
NATO Summit	90,224	88,508

Results on FIFA Worldcup Dataset

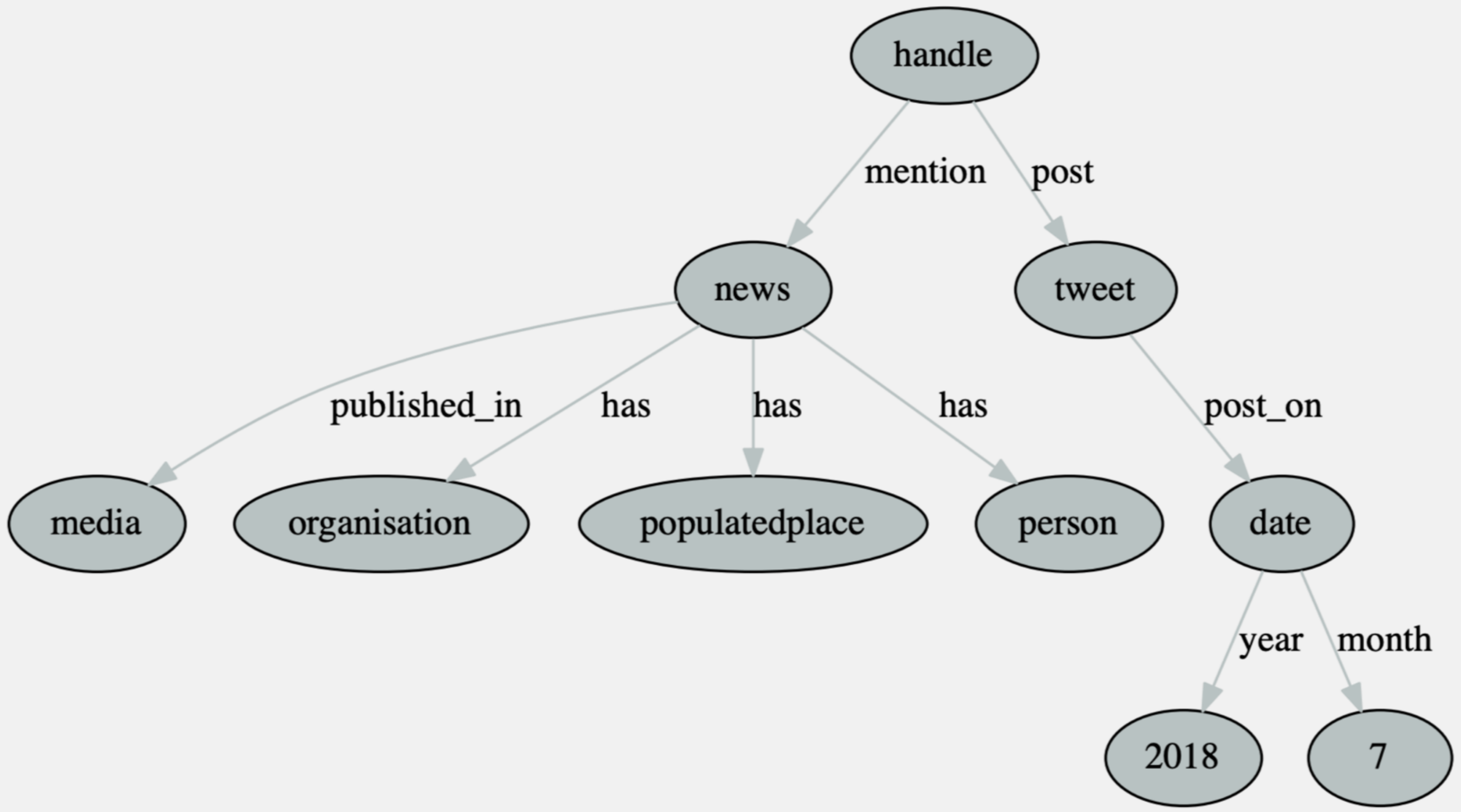


Fig. Normative Pattern on FIFA Worldcup Dataset

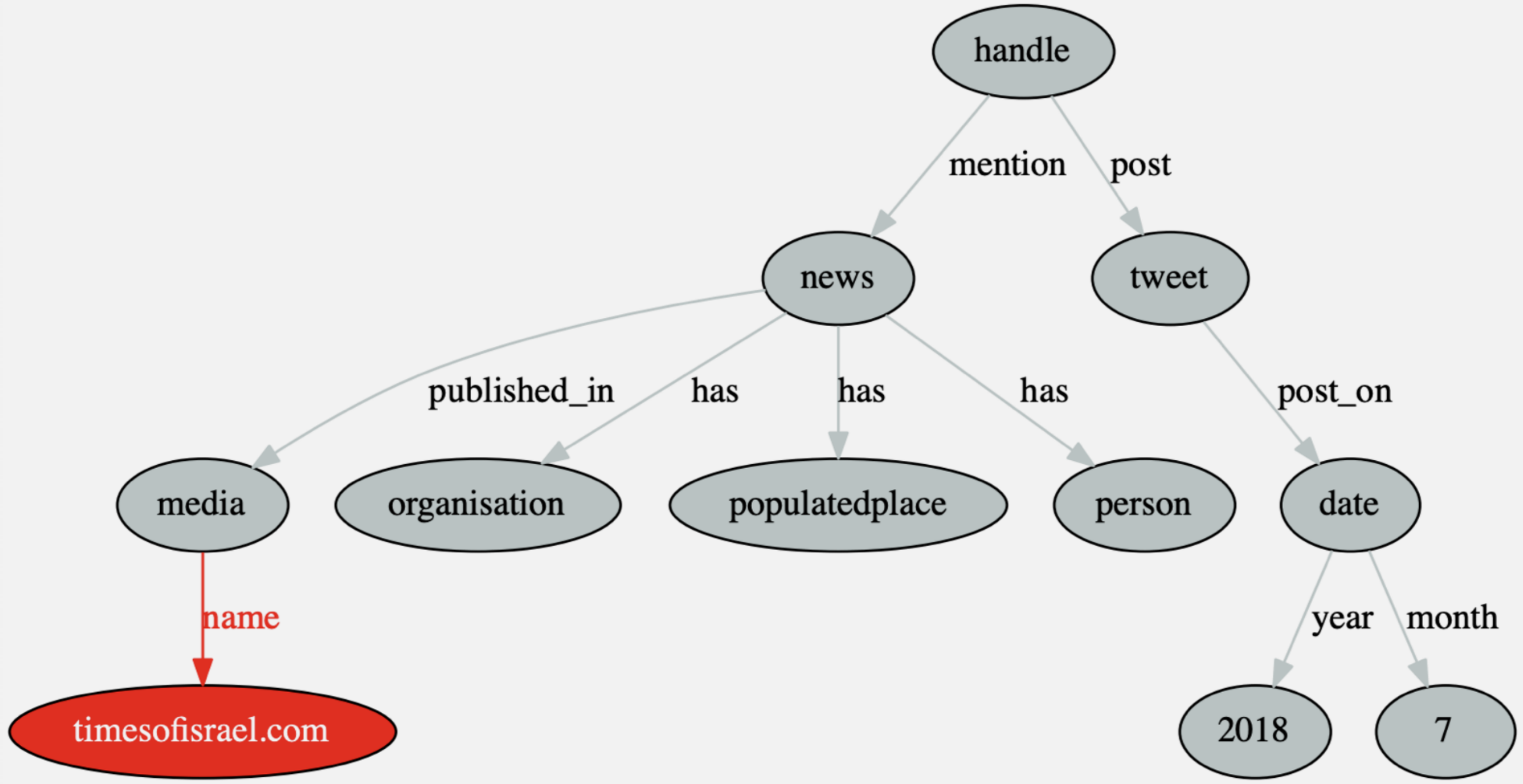
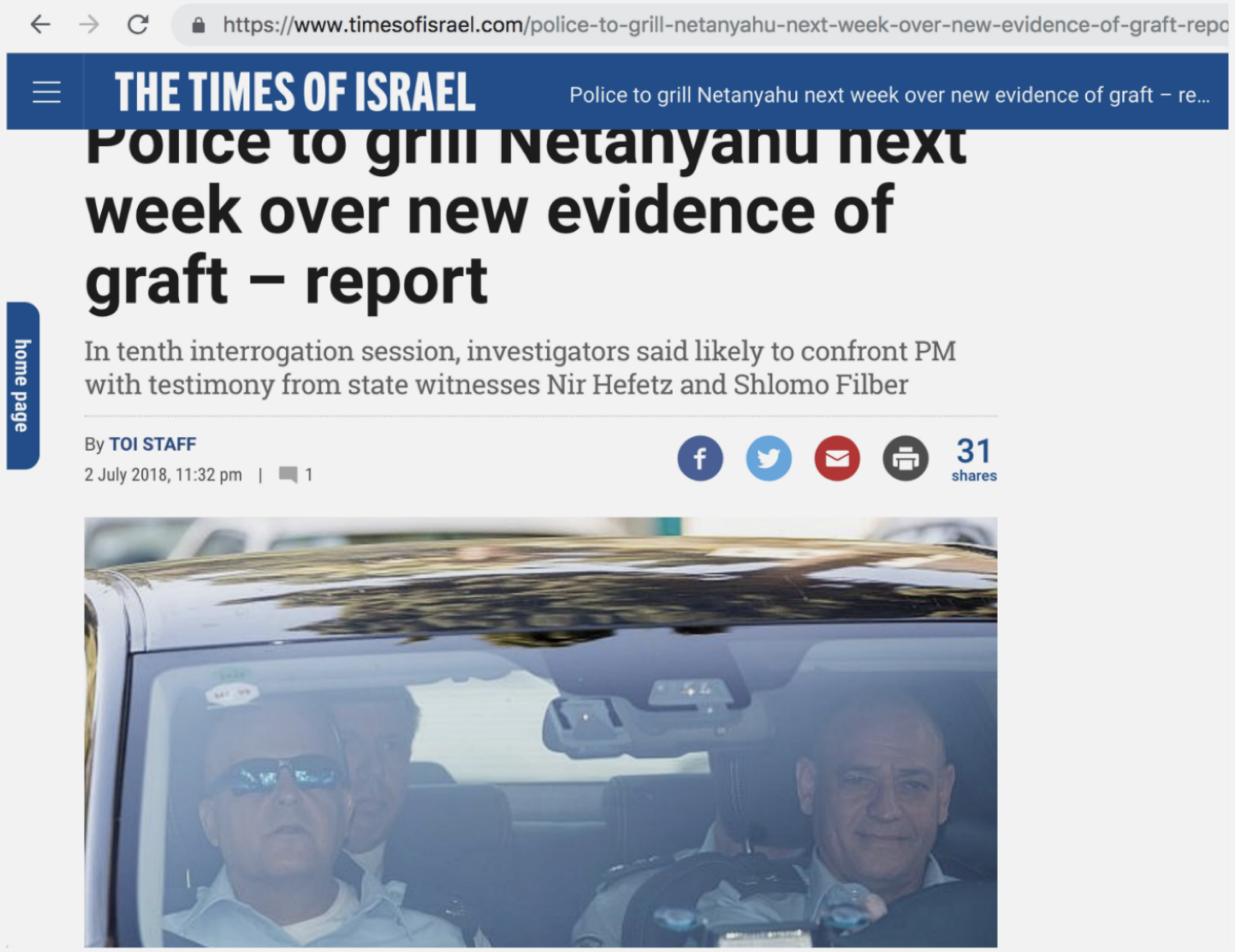


Fig. Anomalous pattern with an additional node

Results on FIFA Worldcup Dataset



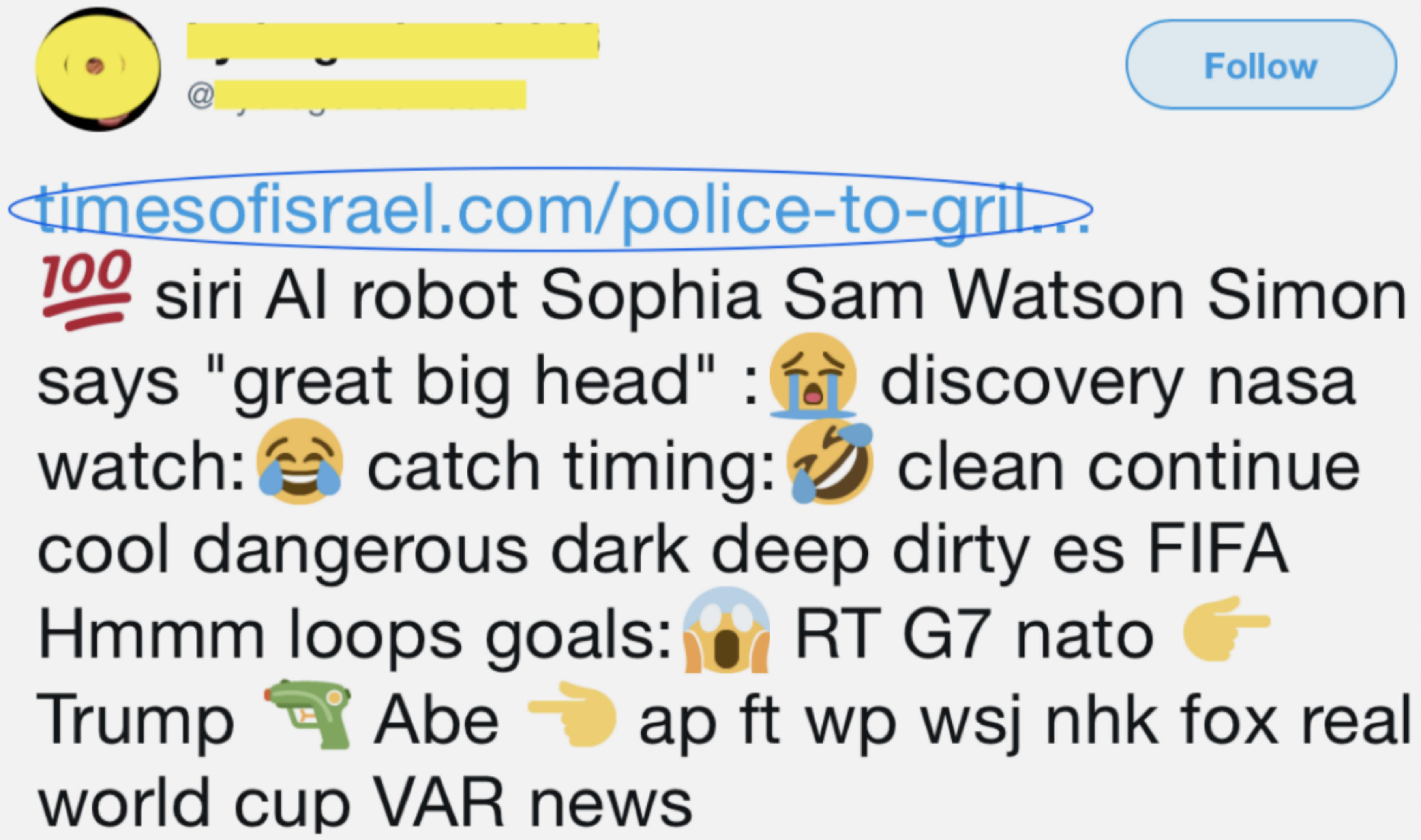


Fig. Example screenshot of a) news and b) tweet showing keyword hijacking

Results on NATO Summit Dataset

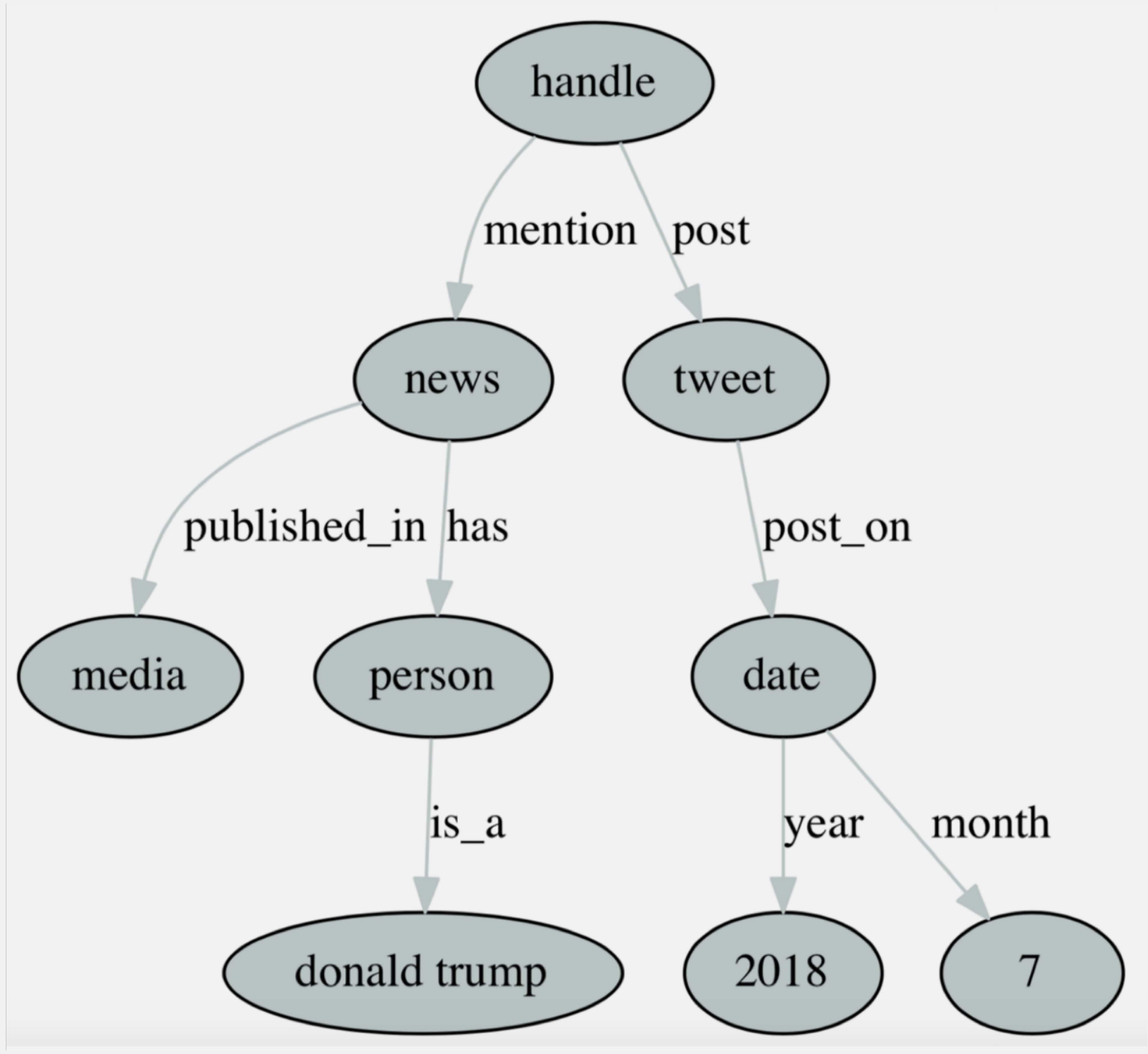


Fig. Normative Pattern on NATO Summit Dataset

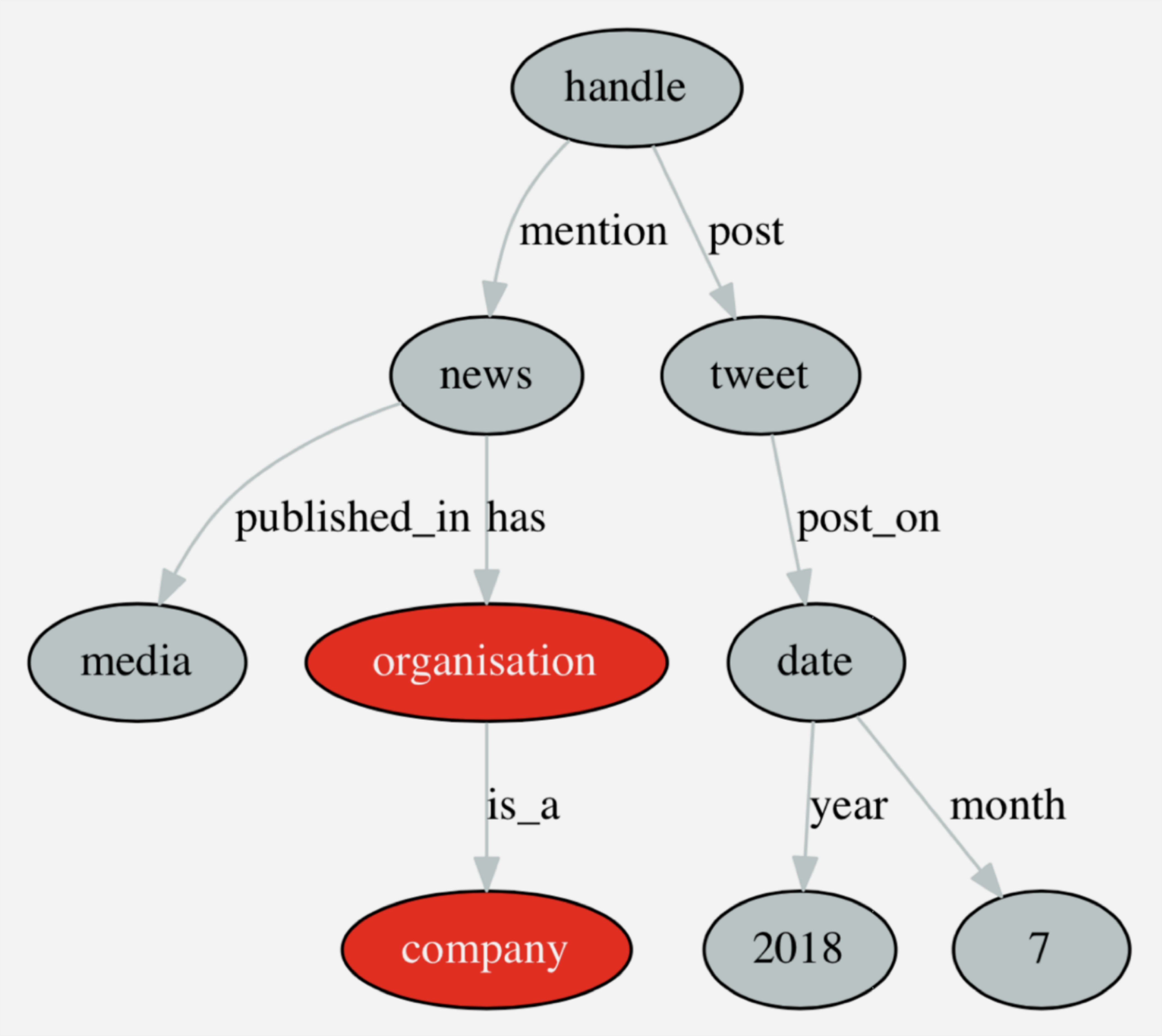


Fig. Anomalous pattern on NATO Summit Dataset

Results on NATO Summit Dataset

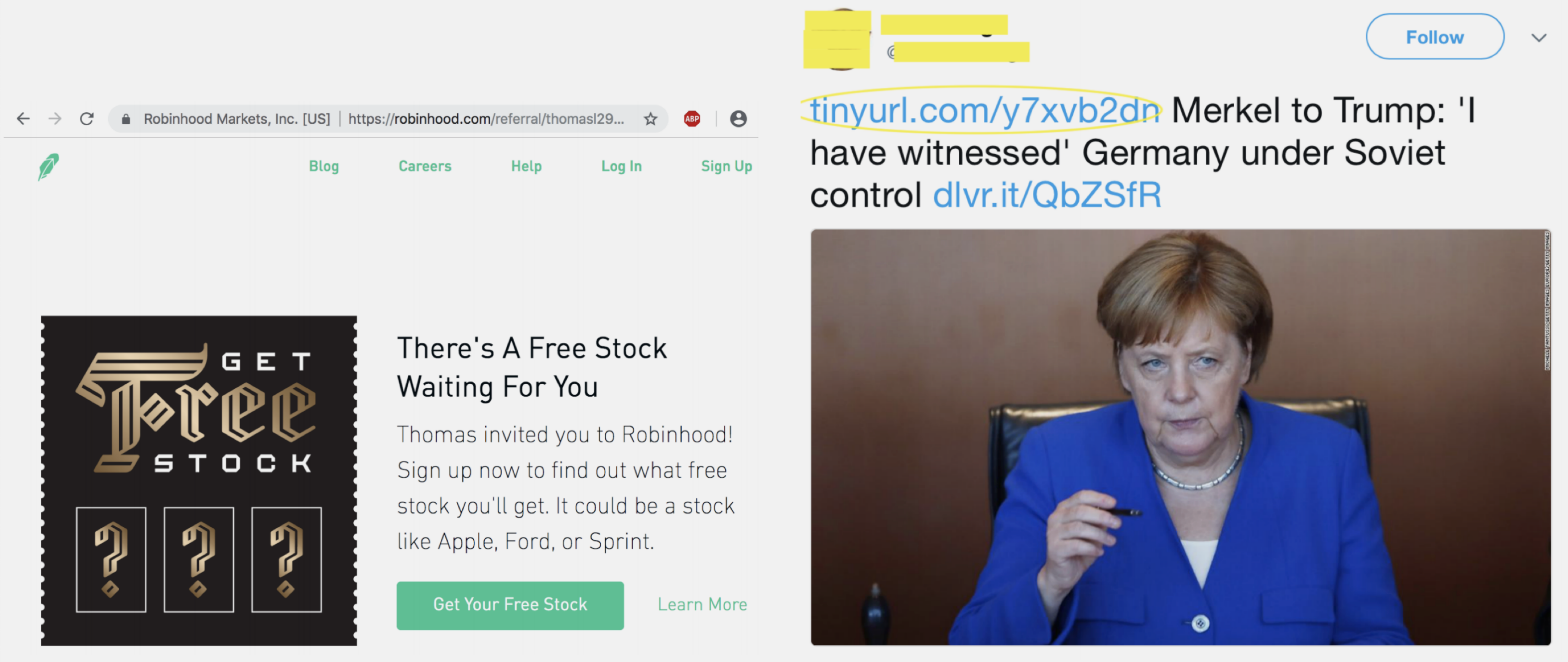


Fig. Example screenshot of a) news and b) tweet showing bogus link anomaly in NATO Summit Dataset

Evaluation

Approach	Precision	Recall	F1-Score
FIFA World Cup Dataset			
Graph-based (non-parametric)	0.15	**1**	0.27
Graph-based (parametric)	0.51	**0.73**	**0.6**
Benevenuto et al. [1]	**0.8**	0.18	0.3
Chen et al. [2]	0.78	0.45	0.58
Anantharam et al.[3]	0.24	0.36	0.29
Boididou et al.[4]	0.7	0.41	0.51
NATO Summit Dataset			
Graph-based(non-parametric)	0.14	1	0.24
Graph-based (parametric)	**0.54**	**0.64**	**0.58**
Benevenuto et al.[1]	0.33	0.02	0.04
Chen et al.[2]	0.52	0.3	0.36
Anantharam et al. [3]	0.24	0.36	0.29
Boididou et al.[4]	0.38	0.27	0.32

- Baseline Approaches (Benevenuto et al., Chen et al., Boididou et al.) suffer from class imbalance problem
- Anantharam et al. needs a predefined reliable information source for each topic
- Proposed graph-based approach is **unsupervised** as well as performs better

Our Advantages

- Does not suffer from class imbalance problems like the baselines
- The performance of proposed approach gets better in data where spam is rare.
- Does not need any prior information as it is completely unsupervised
- Context (named entities and their relationship) generated from both tweet and document (heterogeneous sources) pointed to by URL in tweet is hard to fabricate by the spammer

Conclusion

- Proposed an unsupervised graph-based approach for detecting spam tweets by generating tweet context using
  - Named Entities and their relationships
  - Ontology of named entities

- Proposed approach has superior performance in terms of recall and F1-score to that of existing approaches

Future Work

- Analyze a near real-time feed by converting data stream into graph stream

- Test the robustness of proposed approach by extending it to more topics

References

[1] Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), vol. 6, p. 12 (2010)

[2] Chen, C., Zhang, J., Chen, X., Xiang, Y., Zhou, W.: 6 million spam tweets: A large ground truth for timely twitter spam detection. In: Communications (ICC), 2015 IEEE International Conference on, pp. 7065–7070. IEEE (2015)

[3] Anantharam, P., Thirunarayan, K., Sheth, A.: Topical anomaly detection from twitter stream. In: Proceedings of the 4th Annual ACM Web Science Conference, pp. 11–14. ACM (2012)

[4] Boididou, C., Papadopoulos, S., Apostolidis, L., Kompatsiaris, Y.: Learning to detect misleading content on twitter. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 278–286. ACM (2017)

THANK YOU!

Any QUESTIONS?