Gold Coast DASFAA 2018

Discovering Graph Patterns for Fact Checking in Knowledge Graphs

Peng Lin Qi Song Jialiang Shen Yinghui Wu

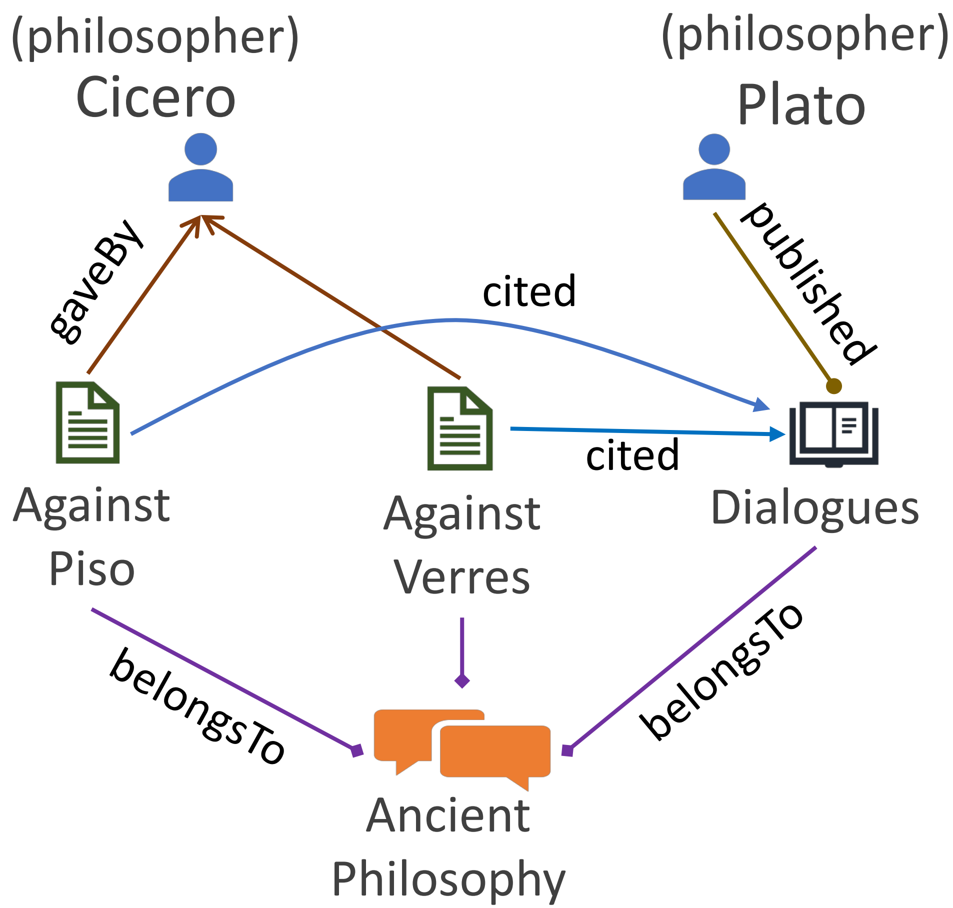
Washington State University

Beijing University of Posts and Telecommunications

Washington State University
Pacific Northwest National Laboratory

What is fact checking?

Knowledge Graph (KG): G=(V, E, L)



Fact: a triple predicate

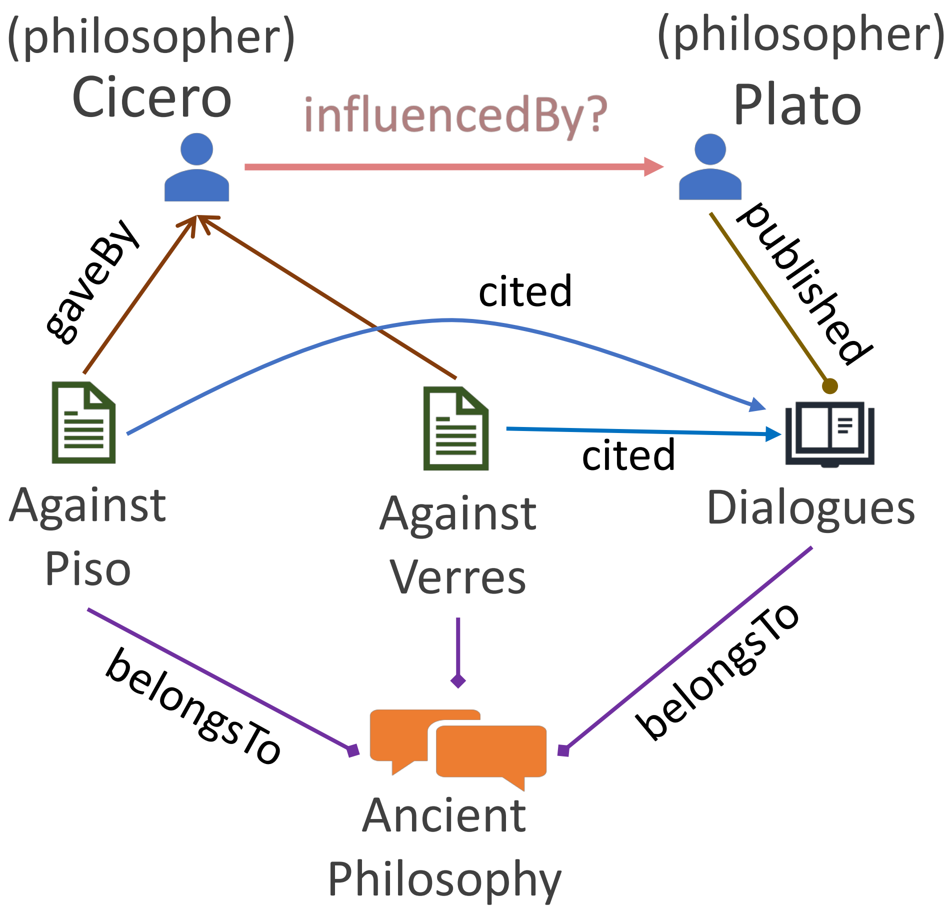
Triple $< v_x, r, v_y >$
- v_x and v_y are two nodes;
- x and y are node labels;
- r is a relationship;

e.g.,
<Cicero, influencedBy, Plato>

- v_x = "Cicero", v_y = "Plato"
- x, y = "philosopher"
- r = "influencedBy"

Fact checking answers if a fact belongs to the missing part of KG.

Fact Checking in Graphs

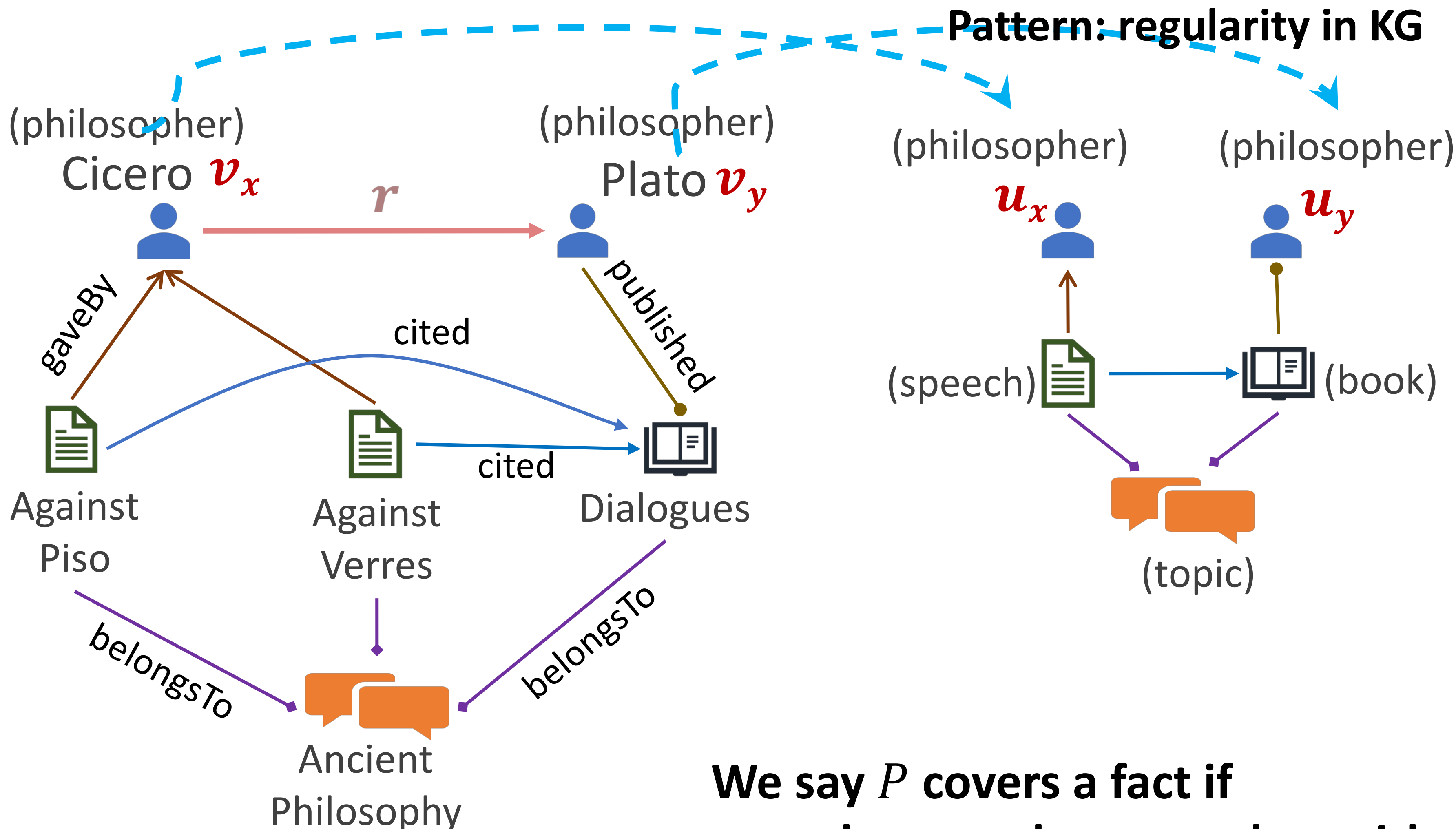


"If a philosopher **X** gave one or more speeches, which cited a book of another philosopher **Y** with the same topic, then the philosopher **X** is likely to be InfluencedBy **Y**."

A fact can be supported by its surrounded substructures!

Graph structure can be evidence for fact checking.

Fact Checking via Graph Patterns



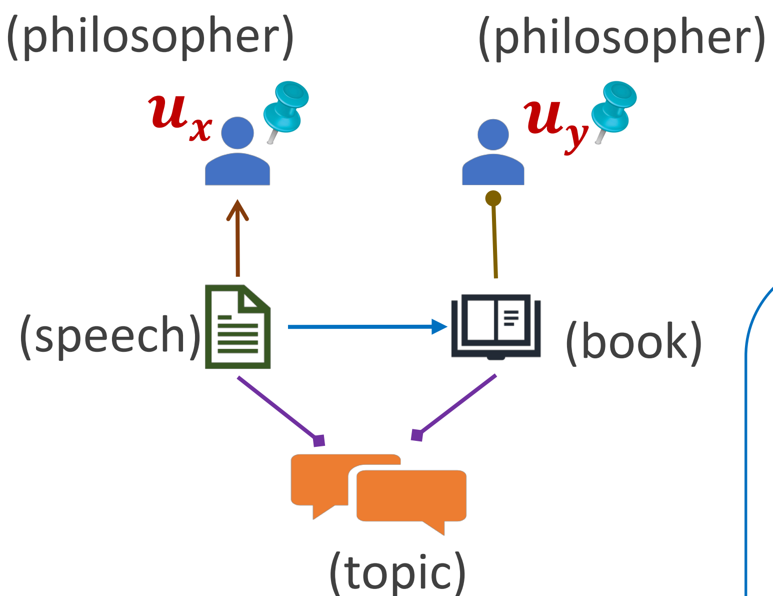
We say P covers a fact if v_x and v_y matches u_x and u_y with r.

Graph structure can be evidence for fact checking.

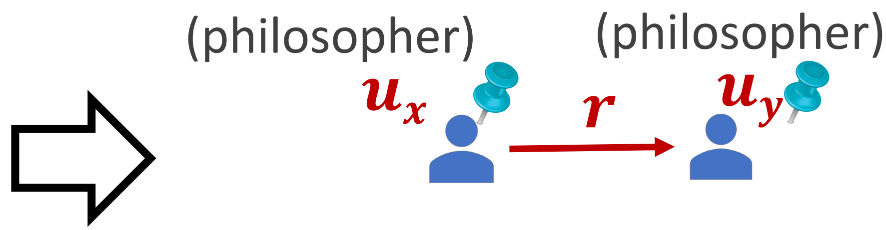
Rule Model: Graph Fact Checking Rules (GFC)

GFC $\varphi : P(x, y) \rightarrow r(x, y)$

LHS



RHS



Rule Semantics:

- GFC φ states that if pattern $P(x, y)$ covers a fact $< v_x, r, v_y >$, then it is true.

Rule matching:

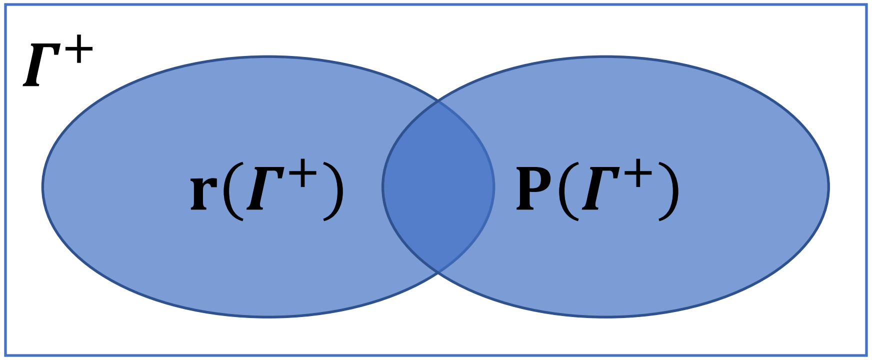
- **Subgraph isomorphism**
  overkill: redundant, too strict, too many
- **Approximate matching**
  (S. Ma, VLDB 2011)

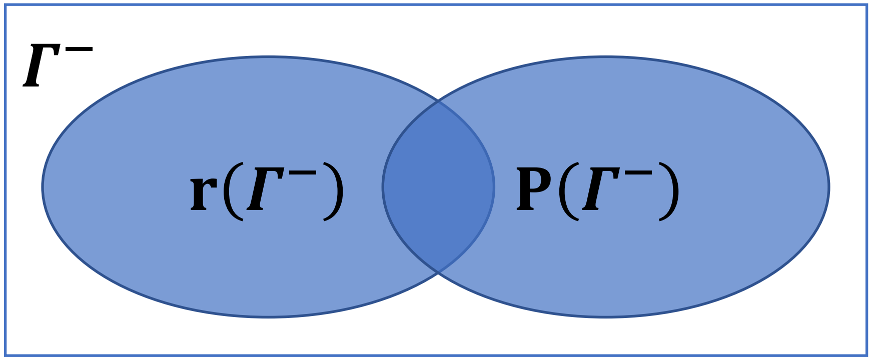
A GFC rule contains two patterns connected by two anchored nodes.

Rule Statistics

- Given: $\mathrm{G} = (\mathrm{V}, \mathrm{E}, \mathrm{L})$
- **GFC** $\varphi : P(x, y) \rightarrow r(x, y)$

- True facts Γ^{+}:
  - sampled from the edges E in G.
- False facts Γ^{-}:
  - sampled from node pairs (v_x, v_y) that have no r between them.
  - following partial closed world assumption (**PCA**)





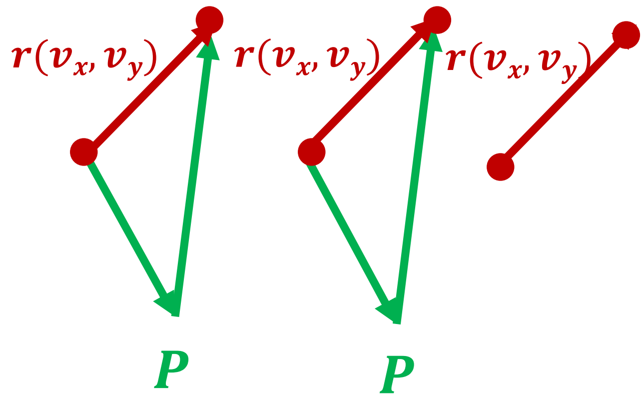
Statistical measures are defined in terms of graph and a set of training facts.

Support and Confidence

GFC: $\varphi : P(x, y) \rightarrow r(x, y)$

- $\text{supp}(\varphi) = \frac{|P(\Gamma^+) \cap r(\Gamma^+)|}{|r(\Gamma^+)|}$

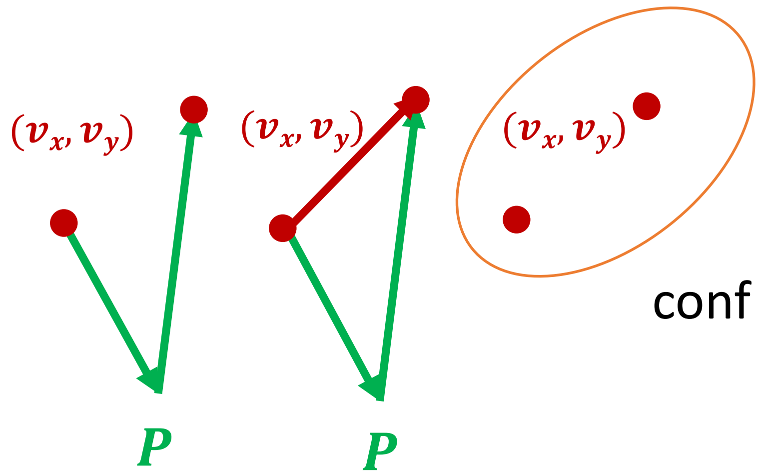
Ratio of facts can be covered out of r(x, y) triples.



supp = 2/3

- $\text{conf}(\varphi) = \frac{|P(\Gamma^+) \cap r(\Gamma^+)|}{|P(\Gamma^+)_N|}$

Ratio of facts can be covered out of (x, y) pairs, under **PCA**.



conf = 1/2

Support and confidence are for pattern mining.

Significance

GFC: $\varphi : P(x, y) \rightarrow r(x, y)$

G-Test score

$$\mathrm{sig}(\varphi, p, n) = 2|\Gamma^{+}|(p \ln \frac{p}{n} + (1 - p) \ln \frac{1 - p}{1 - n})$$

p and n are the supports of $P(x, y)$ for positive and negative facts, respectively.

A "rounded up" score $\max\{\mathrm{sig}(\varphi, p, \delta), \mathrm{sig}(\varphi, \delta, n)\}$ is used in practice.
where δ is a small positive to prevent infinities.

In our work, we also normalize it between 0 and 1 by a sigmoid function.

Significance is the ability to distinguish true and false facts.

Diversity

$\boldsymbol{S}$ is a set of GFCs.

$$\mathrm{div}(\boldsymbol{S}) = \frac{1}{|\Gamma^+|}\sum\nolimits_{t\in\Gamma^+}\sqrt{\sum\nolimits_{\varphi\,\in\,\Phi_t(\boldsymbol{S})}\mathrm{supp}(\varphi)}$$

$\Phi_t(S)$ is the GFCs in S that cover a true fact t.

E.g. $\boldsymbol{S}_1 = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3\}, \boldsymbol{S}_2 = \{\boldsymbol{P}_4, \boldsymbol{P}_5, \boldsymbol{P}_6\}$

	$\boldsymbol{r(vx,vy)_1}$	$\boldsymbol{r(vx,vy)_2}$	$\boldsymbol{r(vx,vy)}$
$\boldsymbol{P}_1$	✓	✓	
$\boldsymbol{P}_2$		✓	✓
$\boldsymbol{P}_3$	✓		✓

	$\boldsymbol{r(vx,vy)_1}$	$\boldsymbol{r(vx,vy)_2}$	$\boldsymbol{r(vx,vy)}$
$\boldsymbol{P}_4$		✓	✓
$\boldsymbol{P}_5$		✓	✓
$\boldsymbol{P}_6$		✓	✓

$\mathrm{div}(\boldsymbol{S}_1)$ = 2 **>** $\mathrm{div}(\boldsymbol{S}_2)$ = 1.6

Diversity is to measure the redundancy of a set of GFCs

Top-$\boldsymbol{k}$ GFC Discovery Problem

To cope with diversity, the total significance $\mathrm{sig}(\boldsymbol{S}) = \sqrt{\sum_{\varphi \in \boldsymbol{S}} \mathrm{sig}(\varphi)}$.

Coverage function: $\mathrm{cov}(\boldsymbol{S}) = \mathrm{sig}(\boldsymbol{S}) + \mathrm{div}(\boldsymbol{S})$

Problem formulation:
Given graph G, support threshold σ and confidence threshold θ, and a set of true facts Γ^+ and a set of false facts Γ^-, and integer k, identify a size-k set of GFCs $\boldsymbol{S}$,
such that:
(a) For each GFC φ in $\boldsymbol{S}$, $\mathrm{supp}(\varphi) \geq \sigma, \mathrm{conf}(\varphi) \geq \theta$.
(b) $\mathrm{cov}(\boldsymbol{S})$ is maximized.

More significance, less redundancy.

Properties of $\mathrm{cov}(\boldsymbol{S})$

- $\mathrm{cov}(\boldsymbol{S})$ is a set function.

  marginal gain: $\mathrm{mg}(\boldsymbol{S}) = \mathrm{cov}(\boldsymbol{S} \cup \{\varphi\}) - \mathrm{cov}(\boldsymbol{S})$

- $\mathrm{cov}(\boldsymbol{S})$ is monotone.

  Adding elements to $\boldsymbol{S}$ does not decrease $\mathrm{cov}(\boldsymbol{S})$.

- $\mathrm{cov}(\boldsymbol{S})$ is submodular.

  If $\boldsymbol{S}_1 \subseteq \boldsymbol{S}_2$ and $\varphi \notin \boldsymbol{S}_2$, then $\mathrm{mg}(\boldsymbol{S}_2) \leq \mathrm{mg}(\boldsymbol{S}_1)$.

Submodularity is a good property for set optimization problem.

Discovery Algorithms

- OPT = $\max\{\mathrm{cov}(\boldsymbol{S})\}$
  - Cannot afford to enumerate every size-k set of GFCs.
  - $\mathrm{cov}(\boldsymbol{S})$ is a monotone submodular function.
  - A greedy algorithm can have $(1-\frac{1}{e})$ approximation of OPT.

- **GFC_batch:**

```
1.  Mine all the patterns satisfying support and confidence.
2.  S = ∅
3.  While |S| < k, do
4.    Select the pattern P with the largest marginal gain.
```

GFC_batch: mining in batch and selecting greedily

Discovery Algorithms

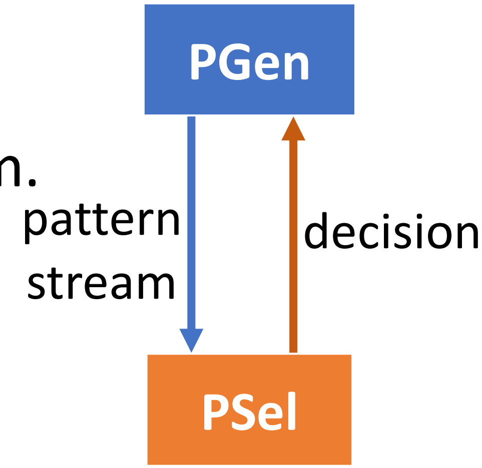
- GFC_batch is infeasible and slow.
  - Still, it requires mine all patterns first.
  - Can we do better?

- **GFC_stream**:
  - Interleave pattern generation and rule selection.
  - Find the top-k GFCs *on-the-fly*.
  - One pass of pattern mining.
  - $(\frac{1}{2} - \epsilon)$ approximation of OPT

GFC_stream: mining and selecting *on-the-fly!*

Discovery Algorithms

- **PGen: pattern generation**
  - Generates patterns *in a **stream** way.*
  - Pass the patterns for selection
  - Can be in any order, e.g., Apriori, DFS, or random.



- **PSel: pattern selection**
  - Selects and constructs GFCs *on-the-fly.*
  - Based on a "sieve" strategy, $\left(\frac{1}{2} - \epsilon\right) \mathrm{OPT}$

Fast compute!

```
1. Estimate the range of OPT by max{cov(P)}
2. Each one is a size-k sieve with an estimation m for OPT.
3. While the sieves are not full
4.      if mg(P,S) ≥ (m/2 − cov(S))/(k − |S|), add P to sieve S.
5. Signal PGen to stop and output the sieve with largest cov.
```

GFC_stream: mining and selecting *on-the-fly!*

GFC-based fact checking

- **GFact$_R$: Using GFCs as rules:**
  - Invokes GFC_stream to find top-k GFCs.
  - "Hit and miss"
    - True if a fact is covered by one GFC.
    - False If no GFC can cover the fact.
  - A typical rule model to compare with: AMIE+

- **GFact: Using GFCs in supervised link prediction:**
  - A feature vector of size k.
  - Each entry encodes the presence of one GFC.
  - Build a classifier, by default, Logistic Regression.
  - A typical rule models to compare with: PRA

Experiment settings

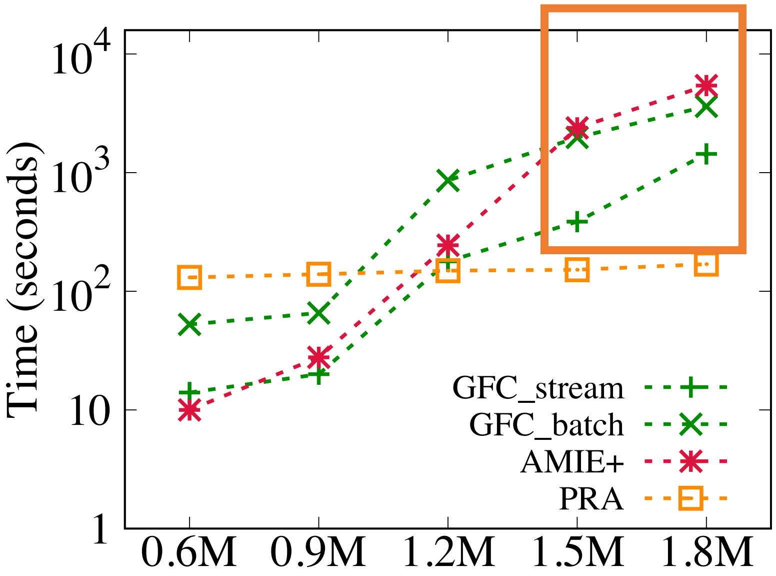
Dataset	category	|V|	|E|	# node labels	# edge labels	# $< x, r, y >$
Yago	Knowledge base	2.1 M	4.0 M	2273	33	15.5 K
DBpedia	Knowledge base	2.2 M	7.4 M	73	584	8240
Wikidata	Knowledge base	10.8 M	41.4 M	18383	693	209 K
MAG	Academic network	0.6 M	1.71 M	8665	6	11742
Offshore	Social network	1.0 M	3.3 M	356	274	633

Tasks	Rule Mining	Fact Checking
Our methods	**GFC_batch**, **GFC_stream**	**GFact**, **GFact$_R$**
Baselines	AMIE+, PRA	AMIE+, PRA, KGMiner
Evaluation Metrics	running time vs. $\|E\|, \|\Gamma^+\|$	prediction rate, precision, recall, F1

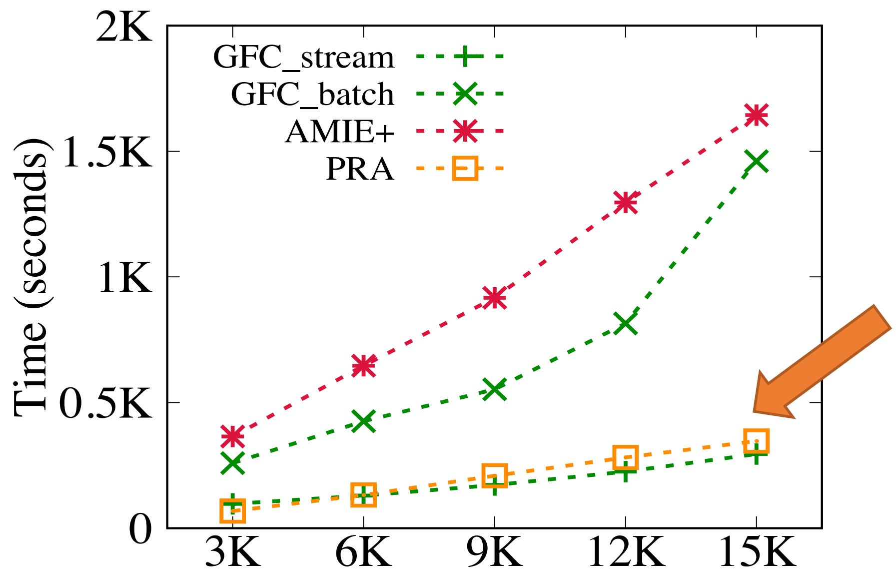
Experiment: efficiency

Overview

- GFC_stream takes 25.7 seconds to discover 200 GFCs over Wikidata with 41.4 million edges and 6000 training facts.
- On average, GFC_stream is 3.2 times faster than AMIE+ over DBpedia.



Varying $|E|$ (DBPedia)

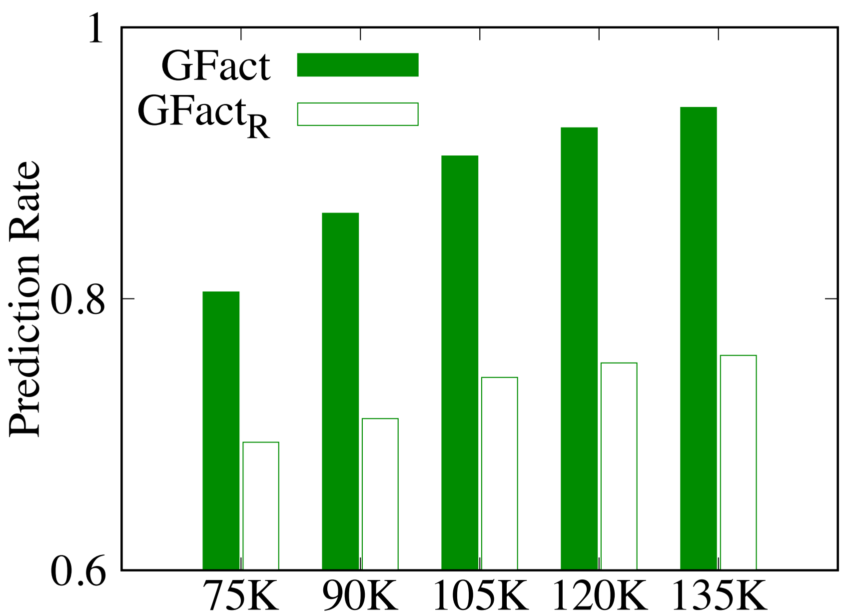


Varying $|\Gamma^+|$ (DBpedia)

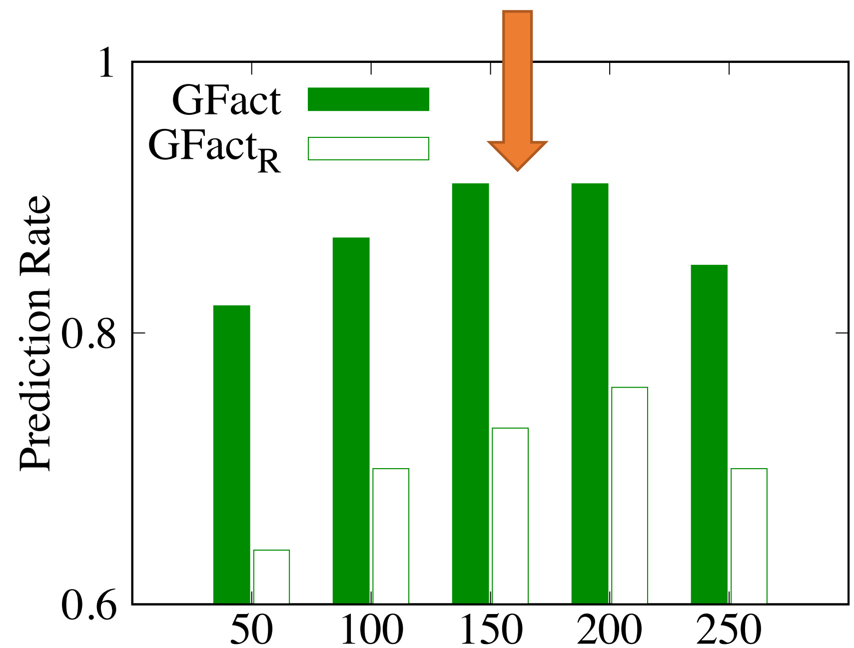
Experiment: effectiveness

Compared with AMIE+, PRA and KGMiner, respectively, on average:

- GFact achieves additional 30%, 20%, and 5% gains of precision over DBpedia.
- GFact achieves additional 20%, 15%, and 16% gains of F1-score over Wikidata.



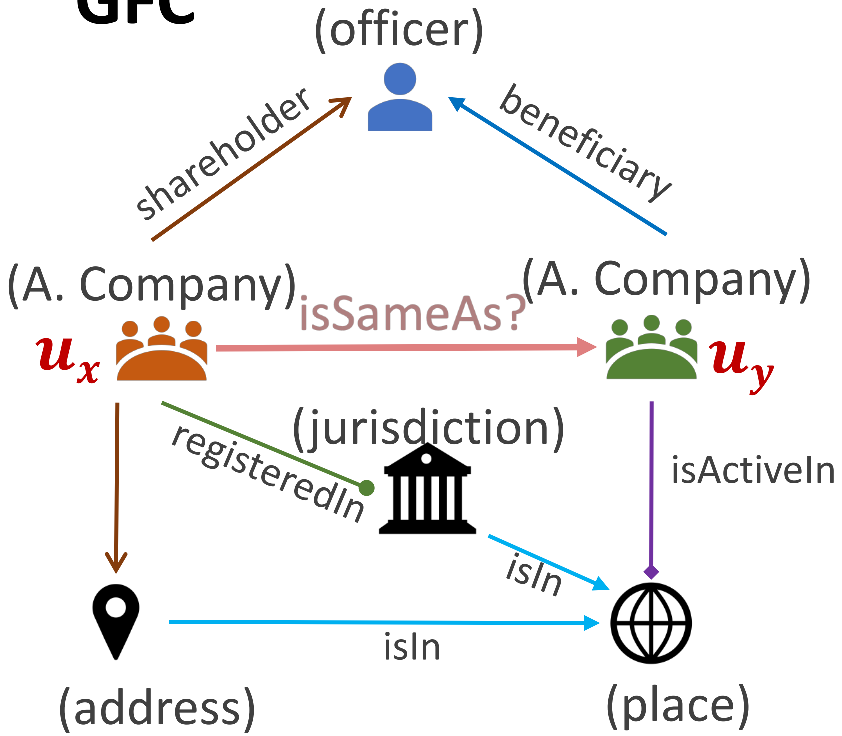
Varying $|\Gamma^+|$ (Wikidata)



Varying k (Wikidata)

Case study: are two anonymous companies same? (Offshore)

GFC



- If an officer is both a shareholder of company u_x and a beneficiary of company u_y, and u_x has an address and is registered through a jurisdiction in a place, and u_y is active in the same place, then they are likely to be the same anonymous company.

AMIE+

- If two anonymous companies are registered in the same place, then they are same.
- Low accuracy.

Conclusions and future work

- *Graph Fact Checking Rules* (GFCs)
- *Top-k GFCs discovery problem*

  Maximize a submodular cov *function.*
- A stream-based rule discovery algorithm
  - One pass, $\left(\frac{1}{2} - \epsilon\right)$ OPT
- Evaluation of GFCs-based techniques
  - Rule models, fact checking (2 methods), efficiency, and case studies.
- Our future work: scalable GFC-based methods
  - Parallel mining, Distributed learning

Sponsored by:

Discovering Graph Patterns for Fact Checking in Knowledge Graphs

Thank you!

Related work: Gstream (IEEE BigData 2017)
Event Pattern Discovery by Keywords in Graph Streams
Mohammad Hossein Namaki, Peng Lin, Yinghui Wu
https://ieeexplore.ieee.org/abstract/document/8258019/