Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket

Tara Safavi (Michigan), Caleb Belth (Michigan), Lukas Faber (ETH Zurich), Davide Mottin (Aarhus), Emmanuel Müller (Bonn-Aachen), Danai Koutra (Michigan)
Knowledge graphs (KGs)

Human-readable representation of facts about the world
Knowledge graphs (KGs)

Question answering, language understanding, recommender systems…

[Gu+ WSDM19, Wang+ KDD19, Logan+ ACL19, …]
Knowledge graphs (KGs)

Millions or billions of nodes and edges, and expanding
Personalized knowledge graph summarization

A new research problem
Personalized knowledge graph summarization
A new research problem

Knowledge graph G

Private, offline, low-resource usage
Security and privacy for personalized, on-device knowledge graphs
Knowledge graphs by definition are enormous, since they aspire to create an entity for every noun in the world, and thus can only reasonably run in the cloud. Realistically, however, most people don’t care about all entities that exist in the world, but rather a small fraction or subset that is personally relevant to them. There is a lot of promise in the area of personalizing knowledge graphs for individual users, perhaps even to the extent that they can shrink to a small enough size to be shippable to mobile devices. This will allow developers to keep providing user value in a privacy-respecting manner by doing more on-device learning and computation, over local small knowledge-graph instances. (We’re eager to collaborate with the research community in pursuit of this goal.)
Security and privacy for personalized, on-device knowledge graphs

Knowledge graphs by definition are enormous, since they aspire to create an entity for every noun in the world, and thus can only reasonably run in the cloud. Realistically, however, most people don’t care about all entities that exist in the world, but rather a small fraction or subset that is personally relevant to them. There is a lot of promise in the area of personalizing knowledge graphs for individual users, perhaps even to the extent that they can shrink to a small enough size to be shippable to mobile devices. This will allow developers to keep providing user value in a privacy-respecting manner by doing more on-device learning and computation, over local small knowledge-graph instances. (We’re eager to collaborate with the research community in pursuit of this goal.)
Our contributions

Personalized knowledge graph summarization: A new research problem

**GLIMPSE**: Efficient framework with performance guarantees via submodular optimization

Large-scale experiments simulating various querying scenarios and user behaviors
Problem 1 (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $Pr(S_u|Q_u)$:

$$\arg\max_{S_u \subseteq G} \log Pr(S_u|Q_u) \text{ s.t. } |T_u| \leq K.$$ (5)
**Problem definition**

**Problem 1** (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $Pr(S_u|Q_u)$:

$$\arg\max_{S_u \subseteq G} \log Pr(S_u|Q_u)$$

$G = (\text{entities } E, \text{relations } R, \text{triples } T)$
Problem 1 (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $\Pr(S_u|Q_u)$:

$$\arg \max_{S_u \subseteq G} \log \Pr(S_u|Q_u)$$

$G = (\text{entities } E, \text{relations } R, \text{triples } T)$
Problem 1 (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $\Pr(S_u | Q_u)$:

$$\arg \max_{S_u \subseteq G} \log \Pr(S_u | Q_u)$$

$G = (\text{entities } E, \text{relations } R, \text{triples } T)$
Problem definition

**Problem 1** (Personalized KG summarization). Given (1) a knowledge graph \( G \), (2) a user \( u \) and her query history \( Q_u \) to \( G \), and (3) a number of triples \( K \), find the personal summary \( S_u = (E_u, R_u, T_u) \subseteq G \) of \( K \) triples that maximizes the log-likelihood of \( \Pr(S_u|Q_u) \):

\[
\arg \max_{S_u \subseteq G} \log \Pr(S_u|Q_u)
\]
Problem definition

**Problem 1** (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $Pr(S_u|Q_u)$:

$$\arg\max_{S_u \subseteq G} \log Pr(S_u|Q_u)$$

A subgraph query to $G$

**Question:** Which English writers influenced William Faulkner?

**Answer:** Charles Dickens, William Shakespeare, …
Problem 1 (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $\Pr(S_u | Q_u)$:

$$\arg\max_{S_u \subseteq G} \log \Pr(S_u | Q_u)$$
Problem definition

**Problem 1** (Personalized KG summarization). Given (1) a knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personalized set $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the likelihood of $\Pr(S_u|Q_u)$:

$$\arg\max_{S_u \subseteq G} \log \Pr(S_u|Q_u) \text{ s.t. } |T_u| \leq K.$$
Problem 1 (Personalized knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $Pr(S_u | Q_u)$:

$$\arg \max_{S_u \subseteq G} \log Pr(S_u | Q_u) \text{ s.t. } |T_u| \leq K.$$  

Construct a "personal summary" $S_u$
Problem 1 (Personalized knowledge graph $G$, (2) a user $u$ and her query history $Q_u$ to $G$, and (3) a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $\Pr(S_u|Q_u)$:

$$\arg \max_{S_u \subseteq G} \log \Pr(S_u|Q_u) \text{ s.t. } |T_u| \leq K.$$
Problem definition

**Problem 1 (Personalized)**

Given a knowledge graph \( G \), a user \( u \) and her query history \( Q_u \), and a number of triples \( K \), find the personal summary \( S_u = (E_u, R_u, T_u) \subseteq G \) of \( K \) triples that maximizes the log-likelihood of \( \Pr(S_u | Q_u) \):

\[
\arg \max_{S_u \subseteq G} \log \Pr(S_u | Q_u) \quad s.t. \quad |T_u| \leq K.
\]

Construct a "personal summary" \( S_u \) given the user's query history \( Q_u \) subject to capacity constraint \( K \)
Problem definition

**Problem 1 (Personalized)**

Given a knowledge graph $G$, a user $u$ and her query history $Q_u$, and a number of triples $K$, find the personal summary $S_u = (E_u, R_u, T_u) \subseteq G$ of $K$ triples that maximizes the log-likelihood of $Pr(S_u|Q_u)$:

$$\arg \max_{S_u \subseteq G} \log Pr(S_u|Q_u) \text{ s.t. } |T_u| \leq K. \quad (5)$$

that maximizes a notion of user-specific "utility" over the knowledge graph $G$. 
GLIMPSE: Graph-based Learning of Personal Summaries

Infer user preferences  ➔  Summary construction

*GLIMPSE is modular: Summary construction step agnostic to how preferences are inferred
Inferring user preferences

Infer user-specific "utility" value for each entity and triple in G
Inferring user preferences

Infer user-specific "utility" value for each entity and triple in G

Entity preference: Interest in a group of facts or "topic"
Inferring user preferences

Infer user-specific "utility" value for each entity and triple in G

**Triple preference**: Interest in a single unit of information

$$
Pr(x_{ijk} | Q_u) \propto Pr(e_i | Q_u)Pr(r_k | Q_u)Pr(e_j | Q_u)
$$

GLIMPSE

Infer user preferences → Summary construction
Summary construction

KG is discrete: Submodular maximization over KG triples
Submodular set functions

\[ \Delta_F(x|A) = F(A \cup \{x\}) - F(A) : \text{marginal gain} \]

\[ F(A \cup \{x\}) - F(A) \]
Submodular set functions

\[ \Delta_F(x|A) = F(A \cup \{x\}) - F(A): \text{marginal gain} \]

\[ F(A) - F(A) \]

**Diminishing returns:** \( F \) is submodular if \( \Delta_F(x|A) \geq \Delta_F(x|B) \)
Submodular set functions

Why do we care? Deep theoretical properties w/ guarantees
Summary construction

KG is discrete: Submodular maximization over KG triples

Intuition: User's "utility" over KG triples has diminishing returns

$$\phi(S_u; Q_u) = \log \Pr(S_u|Q_u) - \log \Pr(S_\alpha)$$

$$= \sum_{e \in E_u} \log \frac{\Pr(e|Q_u)}{\alpha} + \sum_{x_{ijk} \in T_u} \log \frac{\Pr(x_{ijk}|Q_u)}{\alpha}$$

GLIMPSE

Infer user preferences → Summary construction →
Summary construction

KG is discrete: Submodular maximization over KG triples

Intuition: User's "utility" over KG triples has diminishing returns

Show formally that utility is submodular

**Theorem 1.** Equation (7) has a \( (1 - \frac{1}{e}) \)-approximation.

**Proof.** Let \( S_u^{(1)} \subseteq S_u^{(2)} \subseteq G \) be two personal summaries of a knowledge graph \( G \), and let triple \( x_{ijk} = (e_i, r_k, e_j) \in G \setminus S_u^{(2)} \). Consider that (1) if either (or both) entities \( e_i \) or \( e_j \) are contained in \( S_u^{(1)} \), by necessity those entities must...
Summary construction

$(1 - 1/e)$ bound for cardinality-constrained submodular maximization

Greedily select next triple to add to personal summary $S_u$
Summary construction

(1 – 1/e) bound for cardinality-constrained submodular maximization

Greedily select next triple to add to personal summary \( S_u \)

Greedy algorithm is \( O(K|T|) \): Too slow
Optimizing summary construction

OPT1 (personalization): Ignore entities/triples with zero "utility"
Optimizing summary construction

OPT1 (personalization): Ignore entities/triples with zero "utility"

OPT2 (sampling): Sampling-based greedy [Mirzasoleiman+ AAAI15]
Optimizing summary construction

OPT1 (personalization): Ignore entities/triples with zero "utility"

OPT2 (sampling): Sampling-based greedy [Mirzasoleiman+ AAAI15]

OPT3 (lazy updating): Lazy updating of marginal utilities
Optimizing summary construction

Linear in number of triples in G

Theorem 2. GLIMPSE is $O(|T|)$.

Proof. The user preference inference step is linear in $|T|$ using sparse matrix-vector multiplication (Eq. (2)). Then, for the summary construction step, GLIMPSE consists of $K$ iterations, each of which updates $\frac{|T^{\Delta \neq 0}|}{K} \log \frac{1}{\epsilon}$ marginal utilities of sampled triples $x_{ijk} \in A$. Therefore, its runtime complexity is $O(|T| + K \frac{|T^{\Delta \neq 0}|}{K} \log \frac{1}{\epsilon}) = O(|T| + \log \frac{1}{\epsilon} |T^{\Delta \neq 0}|) = O(|T|)$, since $|T^{\Delta \neq 0}| \ll |T|$ and $\log \frac{1}{\epsilon}$ is a constant. □
Optimizing summary construction

Linear in number of triples in G

Theorem 2. GLIMPSE is $O(|T|)$. 

Proof. The user preference inference step is linear in $|T|$ using sparse matrix-vector multiplication (Eq. (2)). Then, for the summary construction step, GLIMPSE consists of $K$ iterations, each of which updates $\frac{|T^\Delta \neq 0|}{K} \log \frac{1}{\epsilon}$ marginal utilities of sampled triples $x_{ijk} \in A$. Therefore, its runtime complexity is $O(|T| + K \frac{|T^\Delta \neq 0|}{K} \log \frac{1}{\epsilon}) = O(|T| + \log \frac{1}{\epsilon} |T^\Delta \neq 0|) = O(|T|)$, since $|T^\Delta \neq 0| \ll |T|$ and $\log \frac{1}{\epsilon}$ is a constant.

GLIMPSE

Tight theoretical guarantees on utility

Theorem 3. GLIMPSE constructs a summary that is a $(1 - \frac{1}{e^{(1-\epsilon)}})$-approximation to the (unknown) optimal personal summary $S_u^*$ for $0 < \epsilon \ll 1$, in expectation.

Proof. As shown in [24], the expected marginal gain of OPT2 for a single triple $x_{ijk}$ is at least

$$E[\Delta_{\phi}(x_{ijk}|S_u;Q_u)] = \frac{1 - \epsilon}{K} \sum_{x_{ijk} \in S_u^* \setminus S_u} \Delta_{\phi}(x_{ijk}|S_u;Q_u), \quad (9)$$

where $S_u^*$ is the (unknown) summary that optimally solves (7). Now, a fact of submodularity, which was proven for $\phi$ in Theorem 1, is that $\sum_{x_{ijk} \in S_u^* \setminus S_u} \Delta_{\phi}(x_{ijk}|S_u;Q_u) \geq \phi(S_u^*;Q_u) - \phi(S_u;Q_u)$, because the sum of individual marginal utilities for each triple must be greater than the total value of those triples grouped as a set, due to diminishing returns. By consequence, combining this fact with the result of (9),

$$E[\Delta_{\phi}(x_{ijk}|S_u;Q_u)] \geq \frac{1 - \epsilon}{K} [\phi(S_u^*;Q_u) - \phi(S_u;Q_u)].$$

Using the above, it can be shown by induction on $K$ that...
**Data**

**TABLE II: Knowledge graphs used in our experiments.**

|        | # Entities \(|E|\) | # Relations \(|R|\) | # Triples \(|T|\) |
|--------|---------------------|---------------------|-------------------|
| DBPedia| 2,026,781           | 1,043               | 10,964,261        |
| YAGO   | 5,155,416           | 72                  | 19,635,755        |
| Freebase| 115,765,760         | 269,984             | 1,000,000,000     |
Data

Real queries to \textit{Freebase}™ [Yih+ ACL16]

Synthetic queries to \textit{yago} [Guu+ EMNLP15, Hamilton+ NeurIPS18]

\begin{table}[h!]
\centering
\caption{Knowledge graphs used in our experiments.}
\begin{tabular}{|l|c|c|c|}
\hline
 & \# Entities $|E|$ & \# Relations $|R|$ & \# Triples $|T|$ \\
\hline
DBPedia & 2026781 & 1043 & 10964261 \\
YAGO & 5155416 & 72 & 19635755 \\
Freebase & 115765760 & 269984 & 1000000000 \\
\hline
\end{tabular}
\end{table}
Data

Topic-based user querying simulation according to literature

[Teevan+ SIGIR07, Radlinski+ ICML08, …]
Data

Topic-based user querying simulation according to literature

In N iterations, sample topic, then sample query from topic

Who influenced Picasso’s art?

[Teven+ SIGIR07, Radlinski+ ICML08, …]
Data

Topic-based user querying simulation according to literature

In N iterations, sample topic, then sample query from topic

[Art] [Pop culture]

Who influenced Picasso’s art?

Where did Picasso die?

[Teevan+ SIGIR07, Radlinski+ ICML08, …]
Data

Topic-based user querying simulation according to literature

In N iterations, sample topic, then sample query from topic

Who influenced Picasso's art?
Where did Picasso die?
Where was Tom Cruise born?

[Tevenan+ SIGIR07, Radlinski+ ICML08, …]
Topic-based user querying simulation according to literature

In N iterations, sample topic, then sample query from topic

Data

[Teevan+ SIGIR07, Radlinski+ ICML08, …]
Data

Topic-based user querying simulation according to literature

In N iterations, sample topic, then sample query from topic

[Art]
- Who influenced Picasso’s art?
- Where did Picasso die?
- Where was Tom Cruise born?
- What inspired Monet?

[Pop culture]
- Who voiced Darth Vader?

[Teevan+ SIGIR07, Radlinski+ ICML08, …]
Query answering task

<table>
<thead>
<tr>
<th>User model</th>
<th>Dataset</th>
<th>TCM</th>
<th>CACHE</th>
<th>PPR-1</th>
<th>PPR-2</th>
<th>PPR-5</th>
<th>PPR-10</th>
<th>GLIMPSE (+ improve.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few topics ((t \in 2 \ldots 5))</td>
<td>DBPedia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many topics ((t \in 5 \ldots 10))</td>
<td>DBPedia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Tang+ SIGMOD16]
## Query answering task

<table>
<thead>
<tr>
<th>User model</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TCM</td>
</tr>
<tr>
<td></td>
<td>CACHE</td>
</tr>
<tr>
<td>Few topics (t \in 2\ldots5)</td>
<td>DBPedia</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
</tr>
<tr>
<td>Many topics (t \in 5\ldots10)</td>
<td>DBPedia</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summarization</th>
<th>Recency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLIMPSE (+ improve.)</td>
<td></td>
</tr>
</tbody>
</table>
## Query answering task

<table>
<thead>
<tr>
<th>User model</th>
<th>Dataset</th>
<th>Summarization</th>
<th>Recency</th>
<th>Personalized PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few topics (t ∈ 2...5)</td>
<td>DBPedia</td>
<td>TCM</td>
<td>CACHE</td>
<td>PPR-1</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td></td>
<td></td>
<td>PPR-2</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td></td>
<td></td>
<td>PPR-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPR-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GLIMPSE (+ improve.)</td>
</tr>
<tr>
<td>Many topics (t ∈ 5...10)</td>
<td>DBPedia</td>
<td>TCM</td>
<td>CACHE</td>
<td>PPR-1</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td></td>
<td></td>
<td>PPR-2</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td></td>
<td></td>
<td>PPR-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PPR-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GLIMPSE (+ improve.)</td>
</tr>
</tbody>
</table>

[Haveliwala+ WWW02, …]
## F1 score comparison

<table>
<thead>
<tr>
<th>User model</th>
<th>Dataset</th>
<th>Summarization</th>
<th>Recency</th>
<th>Personalized PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Few topics ($t \in 2 \ldots 5$)</td>
<td>DBPedia</td>
<td>0.687 ± 0.09</td>
<td>0.684 ± 0.09</td>
<td>0.693 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td>0.539 ± 0.11</td>
<td>0.558 ± 0.10</td>
<td>0.549 ± 0.08</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td>0.678 ± 0.06</td>
<td>0.707 ± 0.05</td>
<td>0.469 ± 0.05</td>
</tr>
<tr>
<td>Many topics ($t \in 5 \ldots 10$)</td>
<td>DBPedia</td>
<td>0.585 ± 0.08</td>
<td>0.603 ± 0.08</td>
<td>0.650 ± 0.08</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
<td>0.526 ± 0.07</td>
<td>0.546 ± 0.07</td>
<td>0.552 ± 0.08</td>
</tr>
<tr>
<td></td>
<td>Freebase</td>
<td>0.542 ± 0.07</td>
<td>0.577 ± 0.05</td>
<td>0.345 ± 0.05</td>
</tr>
</tbody>
</table>

**While baselines perform better or worse under certain conditions, GLIMPSE is most consistent + flexible across user models.**
F1 score comparison

GLIMPSE performs better or on par with best baseline as number of topics vary
GLIMPSE also performs well across various tight resource constraints
Scalability analysis

How important are the optimizations for summary construction?
Scalability analysis

How important are the optimizations for summary construction?

Remove **OPT2+3**: Lazy sampling + updating

<table>
<thead>
<tr>
<th></th>
<th>GLIMPSE</th>
<th>With OPT1 only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Runtime (seconds)</strong></td>
<td>2.11 $\pm$ 0.08</td>
<td>15487.93 $\pm$ 978.05</td>
</tr>
<tr>
<td><strong>Relative to GLIMPSE</strong></td>
<td>1×</td>
<td>7340×</td>
</tr>
</tbody>
</table>
Scalability analysis

How important are the optimizations for summary construction?

Remove OPT1: "Localization" for personalization

<table>
<thead>
<tr>
<th></th>
<th>GLIMPSE</th>
<th>With OPT1 only</th>
<th>With OPT2+3 only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runtime (seconds)</td>
<td>2.11 ± 0.08</td>
<td>15487.93 ± 978.05</td>
<td>28980.46 ± 416.38</td>
</tr>
<tr>
<td>Relative to GLIMPSE</td>
<td>1×</td>
<td>7340×</td>
<td>13734×</td>
</tr>
</tbody>
</table>
Scalability analysis

Theoretically + empirically linear in the number of triples in G
Scalability analysis

Theoretically + empirically linear in the number of triples in $G$

Marginally slower (same order of complexity) but more accurate than baselines
Summary

Personalized knowledge graph summarization:
A new research problem

GLIMPSE: Efficient framework with performance guarantees via submodular optimization

Large-scale experiments simulating various querying scenarios and user behaviors
Work supported by the National Science Foundation, the US Army, Adobe, and Amazon

Thank you + questions
Related work

• Graph summarization and sampling (DB + DM)
  • “One-size-fits-all”, summarizes the whole graph [Liu+ CSUR18]
  • A few recent works consider tasks/users [Amiri+ ICDM18, Kumar and Efstathopoulos VLDB 2019]

• Knowledge graphs (DB + DM + IR + ML + NLP + …)
  • Fact ranking and contextualization [Hasibi+ SIGIR17, Voskarides+ SIGIR18]
  • Complementary to our work

• Personalization and user modeling (DM + IR + ML)
  • Recommendation with additional info from KGs [Zhang+ KDD16, Cao+ WWW19]
  • Not personalized “fact recommendation”
Parameter analysis

Recall: \((1 - 1/e^{(1-\epsilon)})\) approximation for \(\epsilon \ll 1\) in expectation

The higher the parameter \(\epsilon\), the faster but more “approximate” the solution is \(\Rightarrow\) performance drops with \(\epsilon \geq 0.1\)
Conjecture: **Short paths in KGs** more meaningful and useful in tasks → consistent with path-based reasoning, query analysis [Bonifati+ VLDB17, Lin+ EMNLP18]