Examining the Evolution of Identity and Redirection in the LOD Cloud

Shuai Wang

with Idries Nasim, Joe Raad, Peter Bloem, Frank van Harmelen

WAI, 19th September

https://github.com/shuaiwangvu/identity_graph_evolution
https://github.com/shuaiwangvu/redirection

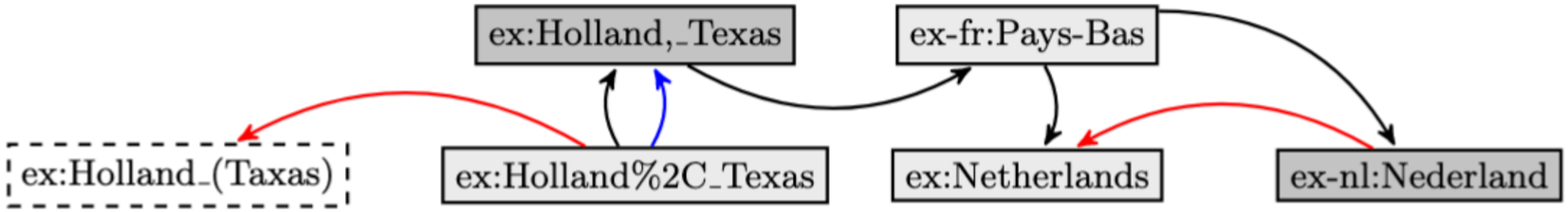
VU VRIJE UNIVERSITEIT AMSTERDAM

OUTLINE

- Introduction and related work
- Evolution of identity graphs
  - Constructing the new identity graphs
  - Compare the identity graphs
- Analysis of redirection
  - Constructing the redirect graphs
  - A qualitative analysis
  - A quantitative analsysis
- Discussion and future work

INTRODUCTION AND RELATED WORK

- Identity crisis [Halpin, et al.]
- 17.9% entities in DBpedia do not exist after 2 years [De Melo, 2013]
- Semantic web evolution [multiple]
- Semantically broken links [Regino et al]



Black = identity links
Red = redirection
Blue = encoding equivalence

RESEARCH QUESTIONS

RQ1: How has the identity graph in the semantic web changed?

RQ2: Can graphs of redirects provide an indication of the evolution of the identity graphs in the semantic web?

RQ3: Can we approximate the implicit semantics of redirection?

RQ4: What are the properties of the redirection graph?

CONSTRUCTING THE NEW IDENTITY GRAPHS

LOD Laundromat in 2015 shows that

- 91.2% of entities are in linksets
- 8.1% of entities are in major hubs with more than 10 identity links.

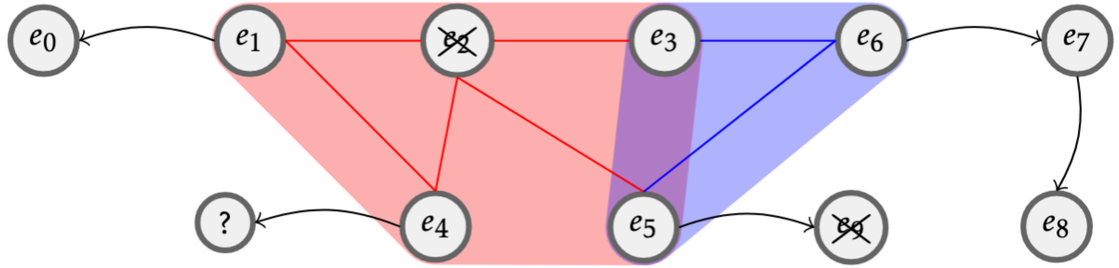
Construct the new identity graph with

- linksets: DBpedia Databus
- Major linked data hub: Yago, Pleiades, WordNet, etc.

Table 1: Sources of the new identity graph

Name/Alias	owl:sameAs	#entities	date of update
DBpedia English links	124.0M	51.3M	Mar 2022
DBpedia-Wikidata links	75.3M	102.5M	Dec 2021
DBpedia external links	61.9M	109.9M	Dec 2021
DBpedia commons links	146.2K	287.3K	Dec 2021
Wikidata	3.7M	6.5K	May 2022
CaLiGraph	8.2M	16.6K	Apr 2021
IMDB	63.2K	92.6K	2020
Yago4	116.3M	183.2M	Mar 2020
Pleiades	117.3K	234.2K	2018
WordNet	117.8K	235.3K	2021
KB	6.3M	12.8M	Jul 2020
GND	15.3M	24.0M	Sep 2021
New identity graph	409.3M	433.4M	Jun 2022

- HTTP 200: ‘OK’
- 400+ HTTP error: Not Found: ‘NF’
- a literal or the request fails: ‘ER’
- Timeout: ‘TO’



- All redirects of 300+
- Redirects Until Found: ‘RUT’
- Redirect Until Not Found: ‘RUNF’
- Redirect Until Error: ‘RUE’
- Redirect Until Found: ‘RUF’

COMPARING THE GRAPHS

	# Triples	#Entities	Size (HDT file)
G (old graph)	558M	179M	4.5G
H (new graph)	409M	443M	11G
I (integrated graph)	951M	555M	15G

57.9M entities are shared

= 32.3% of G and 13.4% of H.

H consists of many more entities than G.

The triple:entity ratio has dropped from 3.12 in G to 0.94 in H, which indicate that redundant edges might be fewer in H.

The HDT file of I is 3.3 times bigger than G.

COMPARING THE GRAPHS

	|Biggest CC|	#CC	Size (HDT file)
G (old graph)	178K	49M	4.5G
H (new graph)	219K	137M	11G
I (integrated graph)	1M	164M	15G

For the largest CC of I.

- 293700 (28%) nodes from G, 450107 (44%) from H, 290877(28%) from both
- 37176(46,5%) CC's from G, 42718(53,4%) CC's from H

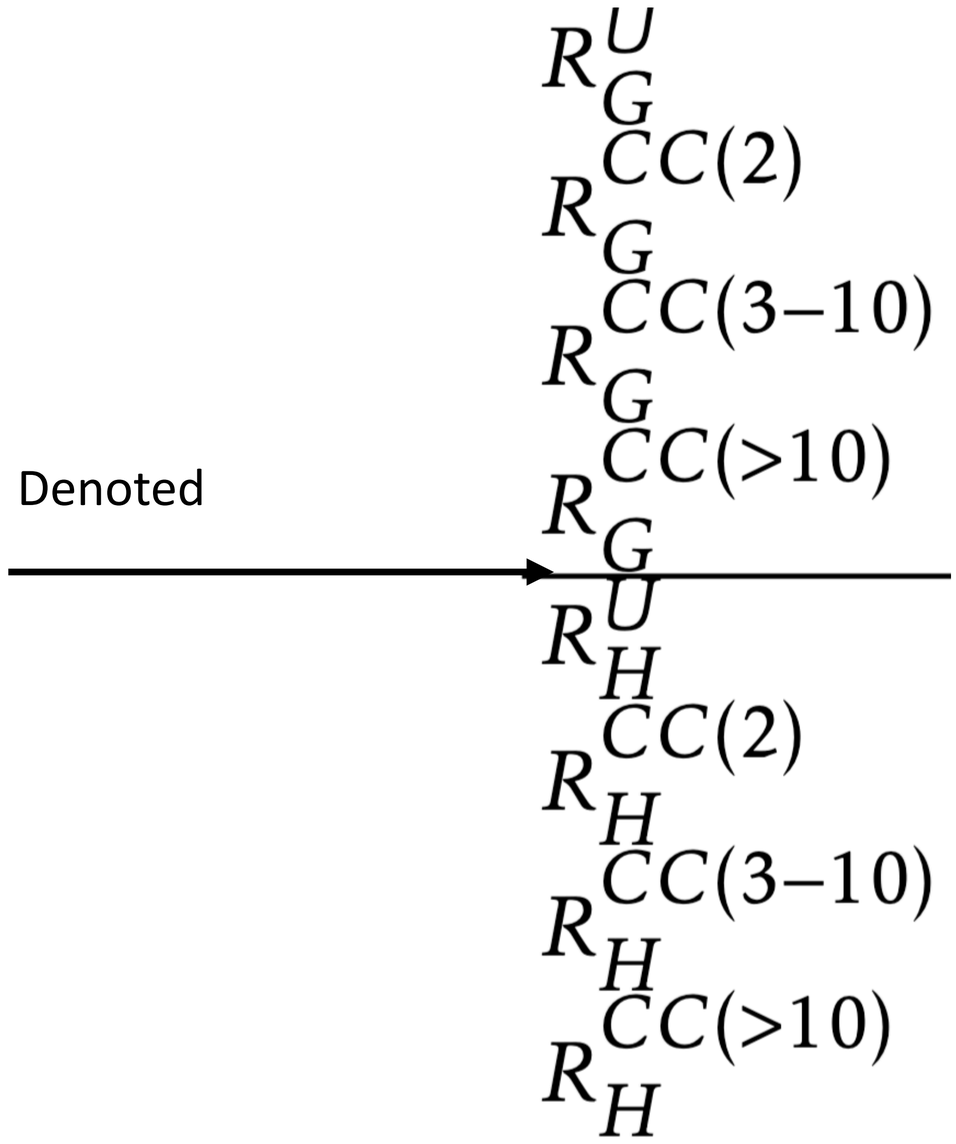
CONSTRUCTING THE REDIRECT GRAPHS

For H and G:

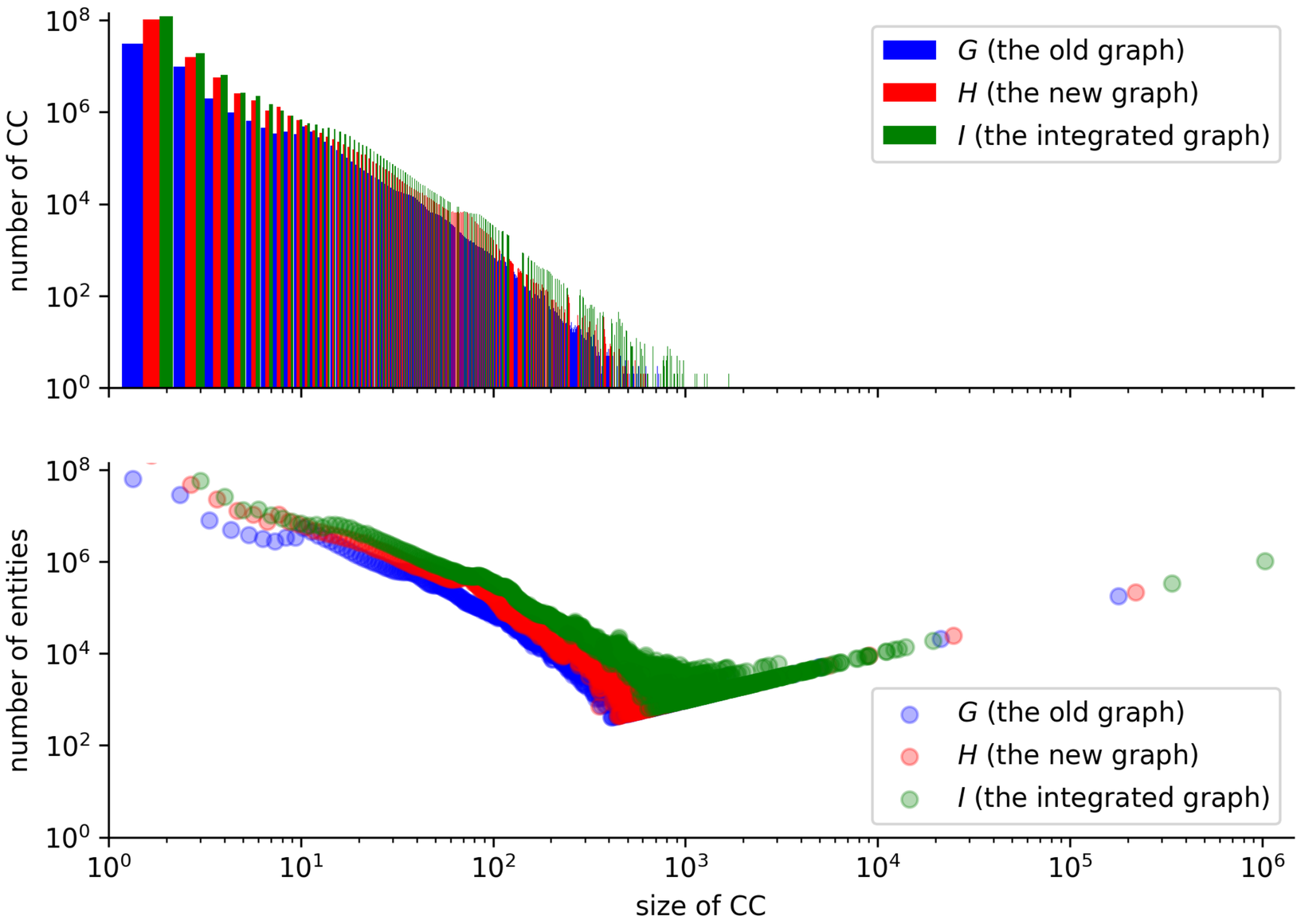
Sampling 100K uniformly

Sampling 20K from

- CC(2)
- CC(3-10)
- CC(>10)



COMPARING THE OLD AND NEW GRAPHS



ANALYSIS OF THE REDIRECTION GRAPH

Table 2: Behavior of HTTP GET request of entities

Graph	NF	OK	ER	TO	RUT	RUNF	RUE	RUF	Valid	Invalid
R_G^U	13.3%	1.1%	23.9%	8.2%	8.1%	12.8%	0.01%	32.6%	33.7%	66.3%
$R_G^{CC(2)}$	4.0%	0.7%	39.5%	12.3%	0.9%	5.5%	0.0%	37.1%	37.8%	62.2%
$R_G^{CC(3-10)}$	8.4%	0.3%	43.4%	5.8%	0.9%	5.8%	5.0%	30.4%	30.7%	69.3%
$R_G^{CC(>10)}$	11.0%	0.8%	26.5%	23.2%	2.3%	10.1%	0.1%	26.0%	26.8%	73.2%
R_H^U	14.8%	3.0%	18.5%	2.4%	4.5%	12.3%	0.1%	44.4%	47.4%	52.6%
$R_H^{CC(2)}$	15.8%	6.7%	3.2%	3.4%	8.0%	8.4%	0.005%	54.5%	61.2%	38.8%
$R_H^{CC(3-10)}$	6.9%	2.4%	59.1%	4.9%	2.5%	6.2%	0.1%	17.9%	20.4%	79.6%
$R_H^{CC(>10)}$	3.4%	4.1%	67.8%	3.9%	2.3%	4.4%	0.05%	14.1%	18.2%	81.8%

Valid = OK + RUF (redirect until found)

Invalid = the rest

ANALYSIS OF THE REDIRECTION GRAPH

Table 2: Behavior of HTTP GET request of entities

Graph	NF	OK	ER	TO	RUT	RUNF	RUE	RUF	Valid	Invalid
R_G^U	13.3%	1.1%	23.9%	8.2%	8.1%	12.8%	0.01%	32.6%	33.7%	66.3%
$R_G^{CC(2)}$	4.0%	0.7%	39.5%	12.3%	0.9%	5.5%	0.0%	37.1%	37.8%	62.2%
$R_G^{CC(3-10)}$	8.4%	0.3%	43.4%	5.8%	0.9%	5.8%	5.0%	30.4%	30.7%	69.3%
$R_G^{CC(>10)}$	11.0%	0.8%	26.5%	23.2%	2.3%	10.1%	0.1%	26.0%	26.8%	73.2%
R_H^U	14.8%	3.0%	18.5%	2.4%	4.5%	12.3%	0.1%	44.4%	47.4%	52.6%
$R_H^{CC(2)}$	15.8%	6.7%	3.2%	3.4%	8.0%	8.4%	0.005%	54.5%	61.2%	38.8%
$R_H^{CC(3-10)}$	6.9%	2.4%	59.1%	4.9%	2.5%	6.2%	0.1%	17.9%	20.4%	79.6%
$R_H^{CC(>10)}$	3.4%	4.1%	67.8%	3.9%	2.3%	4.4%	0.05%	14.1%	18.2%	81.8%

G has more valid entities than H.

Only 1-3% returns meaningful info directly.

>50% has redirection for uniform sampling

VU

ANALYSIS OF THE REDIRECTION GRAPH

Table 2: Behavior of HTTP GET request of entities

Graph	NF	OK	ER	TO	RUT	RUNF	RUE	RUF	Valid	Invalid
R_G^U	13.3%	1.1%	23.9%	8.2%	8.1%	12.8%	0.01%	32.6%	33.7%	66.3%
$R_G^{CC(2)}$	4.0%	0.7%	39.5%	12.3%	0.9%	5.5%	0.0%	37.1%	37.8%	62.2%
$R_G^{CC(3-10}$	8.4%	0.3%	43.4%	5.8%	0.9%	5.8%	5.0%	30.4%	30.7%	69.3%
$R_G^{CC(>10)}$	11.0%	0.8%	26.5%	23.2%	2.3%	10.1%	0.1%	26.0%	26.8%	73.2%
R_H^U	14.8%	3.0%	18.5%	2.4%	4.5%	12.3%	0.1%	44.4%	47.4%	52.6%
$R_H^{CC(2)}$	15.8%	6.7%	3.2%	3.4%	8.0%	8.4%	.005%	54.5%	61.2%	38.8%
$R_H^{CC(3-10}$	6.9%	2.4%	59.1%	4.9%	2.5%	6.2%	0.1%	17.9%	20.4%	79.6%
$R_H^{CC(>10)}$	3.4%	4.1%	67.8%	3.9%	2.3%	4.4%	0.05%	14.1%	18.2%	81.8%

#Valid decreases as the size of CCs increase, especially H.

Opposite trend for NF, TO, RUNF, RUE

Different for OK is too small to draw a conclusion.

ANALYSIS OF THE REDIRECTION GRAPH G

From now on, the results are only about the old graph.

Table 2: Properties of the redirection graph

Graph	#Entities	#Entities Redirected	#Nodes	#Edges	#Hops	Longest Path (#Hops)
R^{U}	100K	53,487 (53.49%)	169,021	116,031	1.71	8
$R^{CC(2)}$	20K	8,693 (43.46%)	30,091	21,602	1.64	8
$R^{CC(3-10)}$	20K	8,412 (42.06%)	29,697	21,490	-	-
$R^{CC(>10)}$	20K	7,704 (38.52%)	24,914	18,102	2.05	8

LONG REDIRECTION PATHS

['http://dbpedia.org/resource/Mirage_%28pop_group%29',
'http://dbpedia.org/resource/Mirage_(pop_group)',
'https://dbpedia.org/resource/Mirage_(pop_group)',
'http://dbpedia.org/page/Mirage_(pop_group)',
'https://dbpedia.org/page/Mirage_(pop_group)',
'http://dbpedia.org/resource/Mirage_(disambiguation)',
'https://dbpedia.org/resource/Mirage_(disambiguation)',
'http://dbpedia.org/page/Mirage_(disambiguation)',
'https://dbpedia.org/page/Mirage_(disambiguation)']

IMPLICIT SEMANTICS OF REDIRECTION (4,000 EDGES)

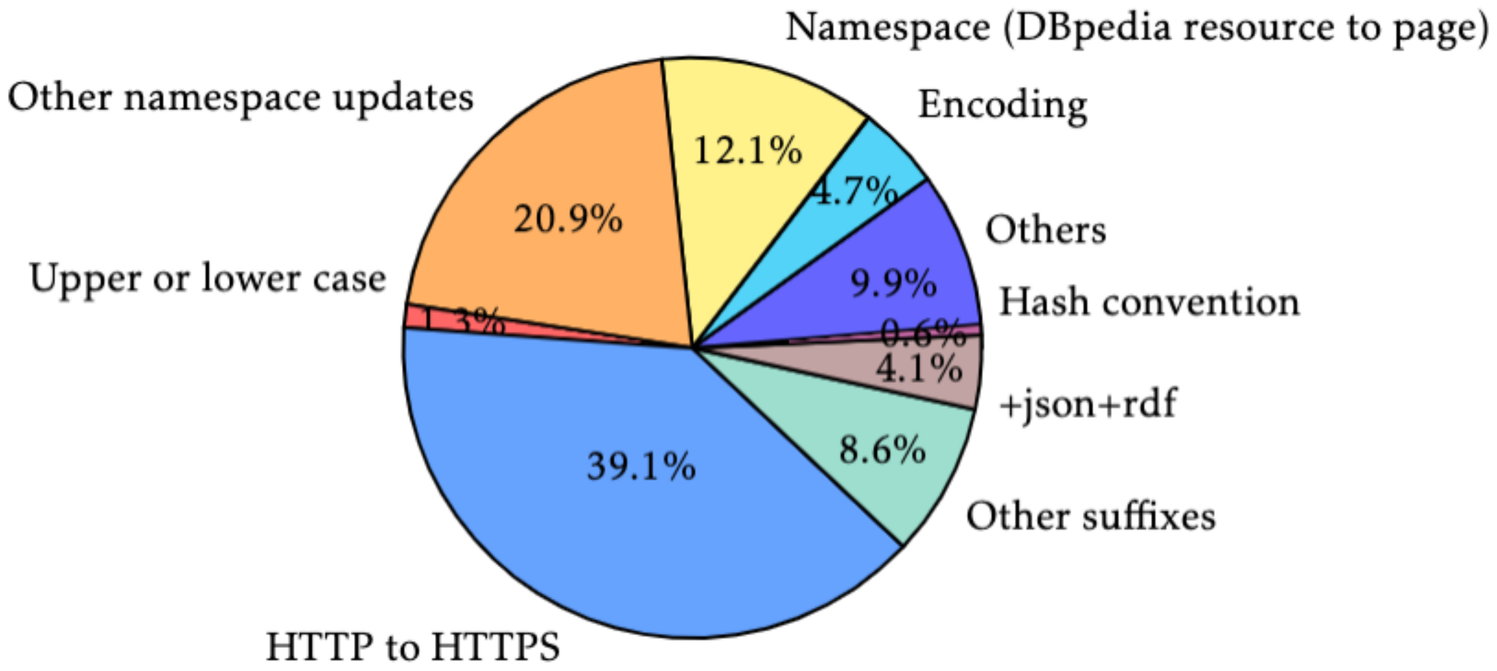


Fig. 2: Proportion of redirection behavior among sampled entities

- 45.1% (encoding, http->https, upper/lower case)
- 16.8% (DBpedia resource to page, etc.)

Approx. 45.1% - 83.2% can be taken as identity links

100 CHAINS OF REDIRECTION

- On average 1.7 hops. We examine redirection chains with over 2 hops
- Similar number of hops for RUF, RUE, RUNF, RUT. So we sample uniformly
- 85% happens within a domain
- Wikidata (28%) and DBpedia (25%) are among the most observed
- Chains of DBpedia are often among the longest.
- Some others were observed for bibsonomy.org (5%) and via.org (1%)

DISCUSSION AND FUTURE WORK

- Few entities can be redirected from G to H
- Redirection is a well-observed in identity graphs
- When only 1-3% can be dereferenced, it hurts accessibility and interoperability
- We observed some correlation between size of CC and the hops of redirect
- HTTPS is well adopted in the semantic web community
- Why we have opposite trend for NF, TO, RUNF, RUE?

VU