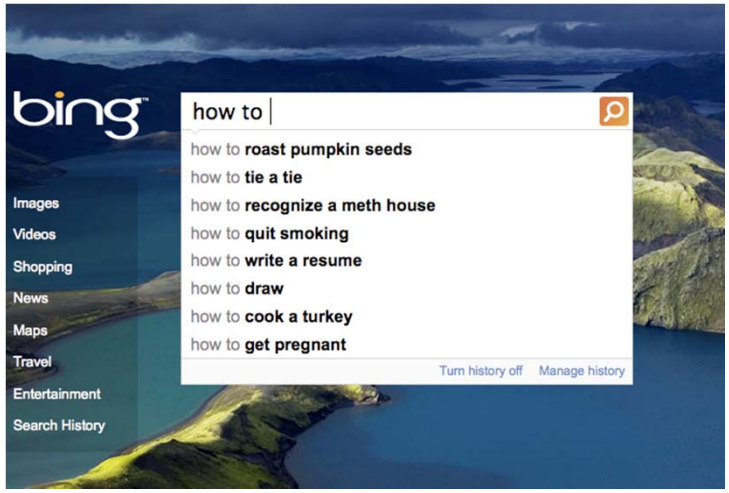
On Matching Web-scale Entity Graphs

Seung-won Hwang
Associate Professor
Department of Computer Science and Engineering
POSTECH, Korea

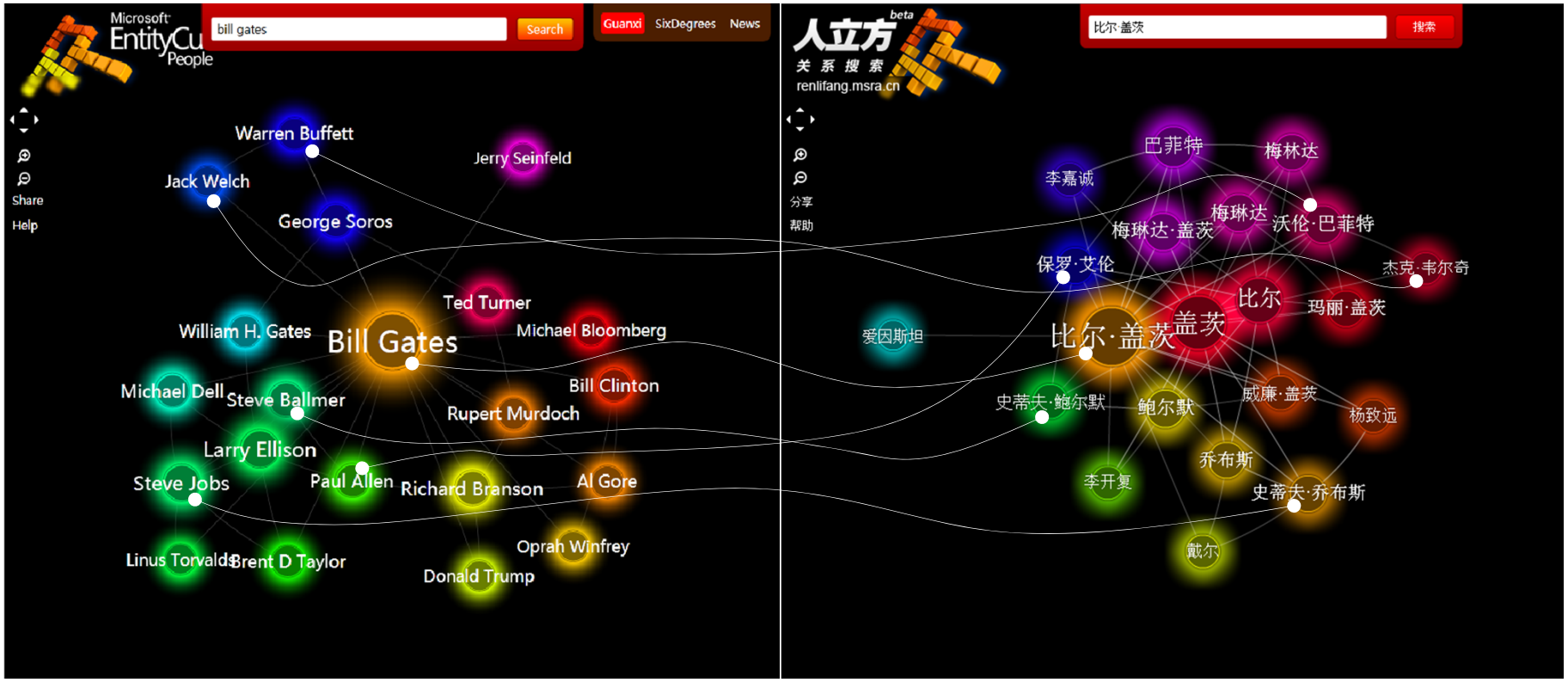
My Research Mantra

- Web is overloaded with human intelligence trails
  - e.g., search logs, clicks
- We can mimic social human intelligence by mining such trails
  - e.g., autoComplete



- This talk focuses on trails represented in a form of entity graphs (finer-granularity)

Motivation



English
EntityCube

Chinese
EntityCube
(a.k.a, Renlifang)

Mining Name Translation [CIKM11,EMNLP10]

Joint work with Gae-won You[1], Jinhan Kim[1]
Ming Zhou[2], Young-In Song[2], Long Jiang[2], Zaiqing Nie[2]

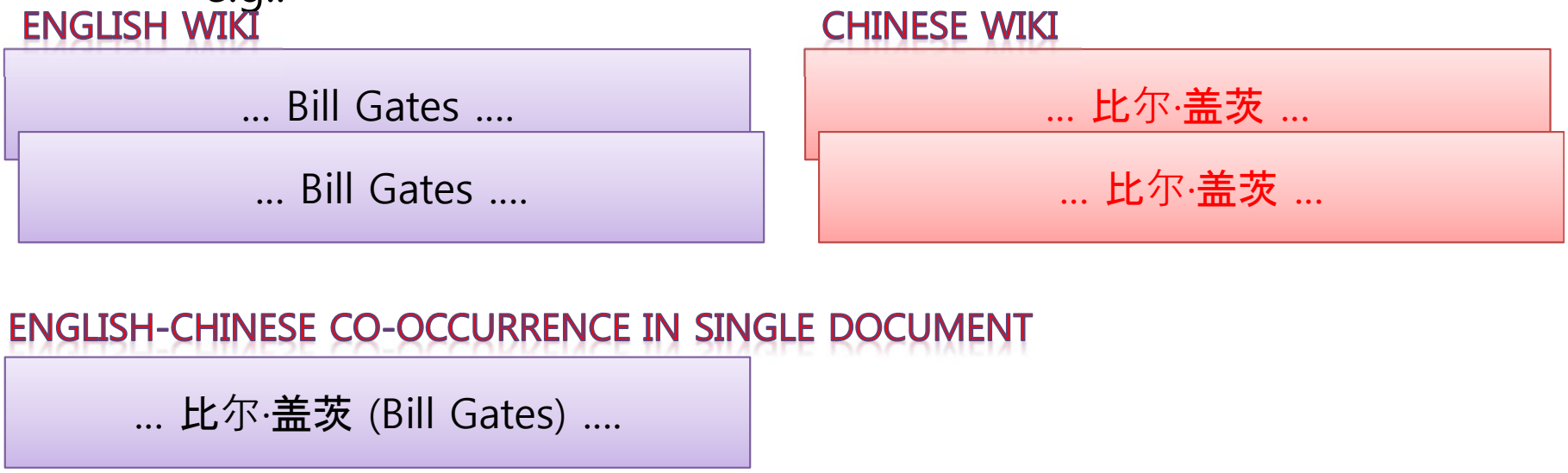
[1]Pohang University of Science and Technology (POSTECH), Korea
[2]Microsoft Research Asia (MSRA), China

(turns out) A novel approach for MT

- Name translation: Hard problem in machine translation (MT)
  - **Transliteration-based**: low accuracy, 100% coverage

    e.g., arbitrary mapping, e.g., "Jackie Chan" to "**成**龙 (chénglóng)"
  - **Corpus-based**: low coverage, high accuracy

    e.g.:



Basic Insight

- We combine the strength of the two approaches

Phonetic similarity

If two names are phonetically similar, they are likely to be a name translation pair

- E.g., "Bill Gates" vs. "比尔·盖茨 "

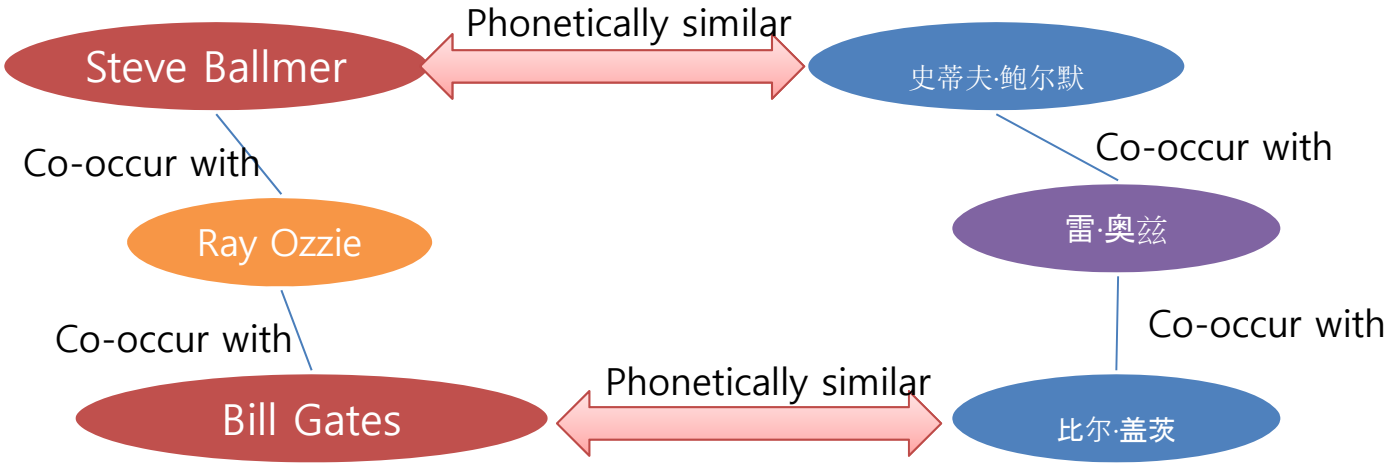
Relational similarity

If two names in source language are related, their corresponding names are also likely to be related in target language

- E.g., ("Jackie Chan", "Bill Gates") vs. ("成龙", "比尔·盖茨 ")

Basic Insights

- Relax using bilingual co-occurrence to monolingual co-occurrence based on **relationa l similarity**



Neighbors of two entities are si milar!
- Relationally similar

Step 1: Initialization of Similarity Matrix

- Denote
  - N_e: Word sequence of given English name,
  - PY_c: Pinyin sequence of Chinese name N_c
- Transliteration similarity can be defined as the edit-distan ce to measure their spelling similarity:

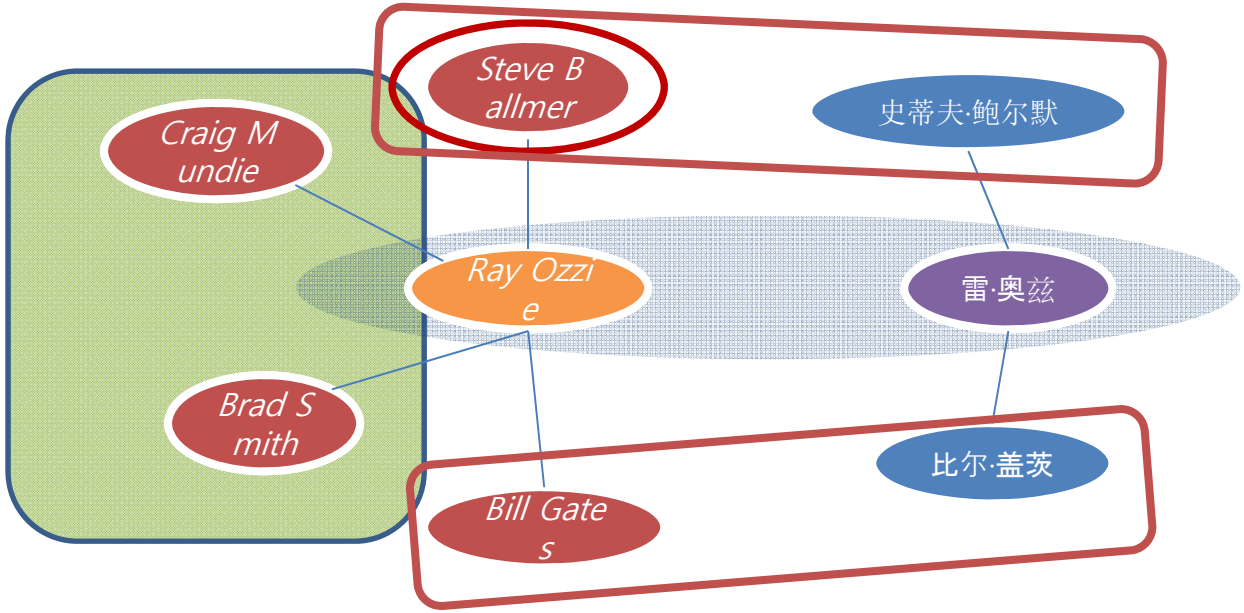
$$Sim_{tl}(N_c, N_e) = 1 - \frac{ED_{name}(N_e, PY_c)}{Len(PY_c) + Len(N_e)}$$

Step 2: Reinforcement

- Update translation similarity between two entities leveraging the relational similarities

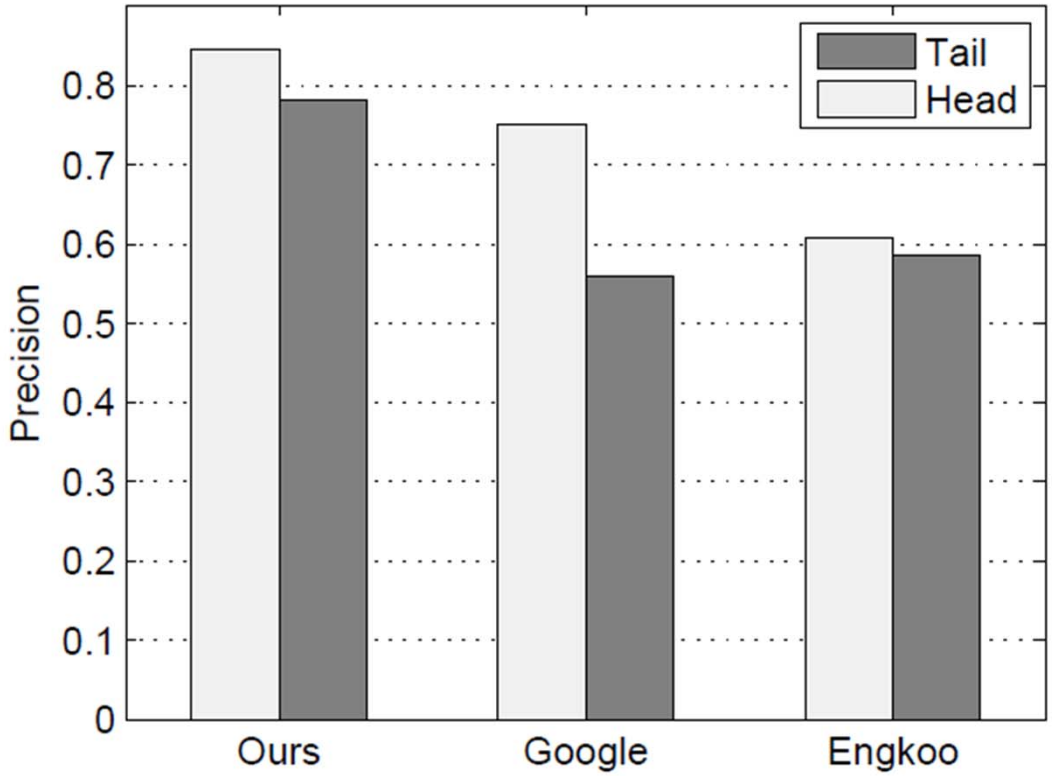
2. Not assigning a similarity between two nodes of the translation pairs *e can be non-matched neighbor pairs*

- Align neighbors of two nodes with similarity scores
  - Consider only the neighbor alignment pairs whose similarity > Θ



Results – Head/Tail Entities [EMNLP10]

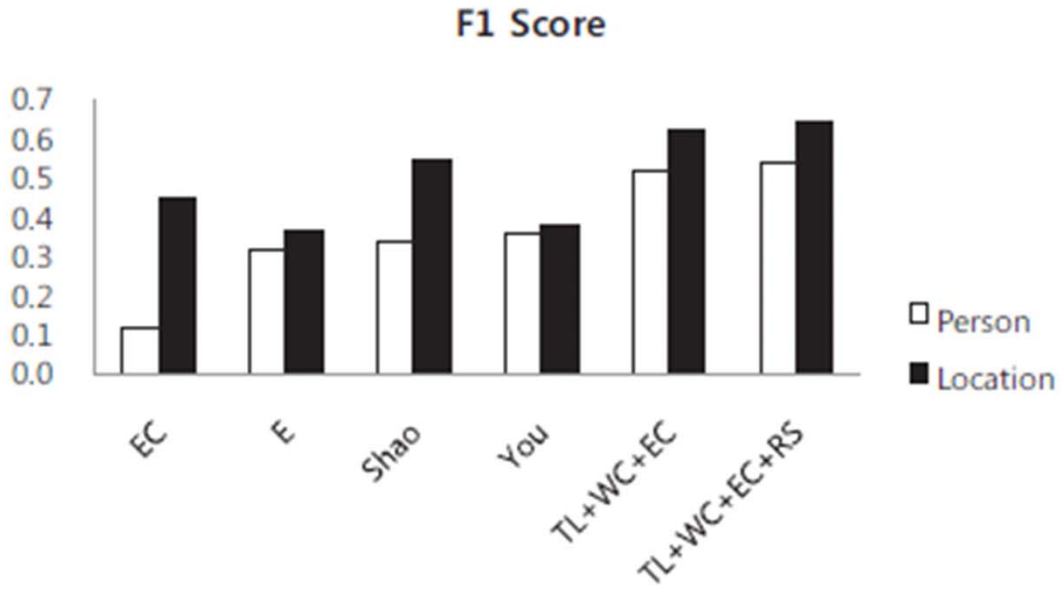
- Divide head and tail entities according to # hits in Google; head (top-100 popular) and tail (remaining)



Missing Quadrant

	Entity	Relationship
Using names	贝拉克·奥巴马~ Barack Obama	(贝拉克·奥巴马, 米歇尔奥巴马)~(Barack Obama, Michelle Obama)
Using context (surrounding words)	总统~ president	配偶~spouse

A Holistic Approach [CIKM11]



Mining Social Network Accounts [EDBT11, SIGKDD09]

Joint work with Jongwuk Lee[1], Gae-won You[1]
Ji-rong Wen[2], Zaiqing Nie[2]

[1]Pohang University of Science and Technology (POSTECH), Korea
[2]Microsoft Research Asia (MSRA), China

EntityCube and Twitter Network

ceoSteveJobs

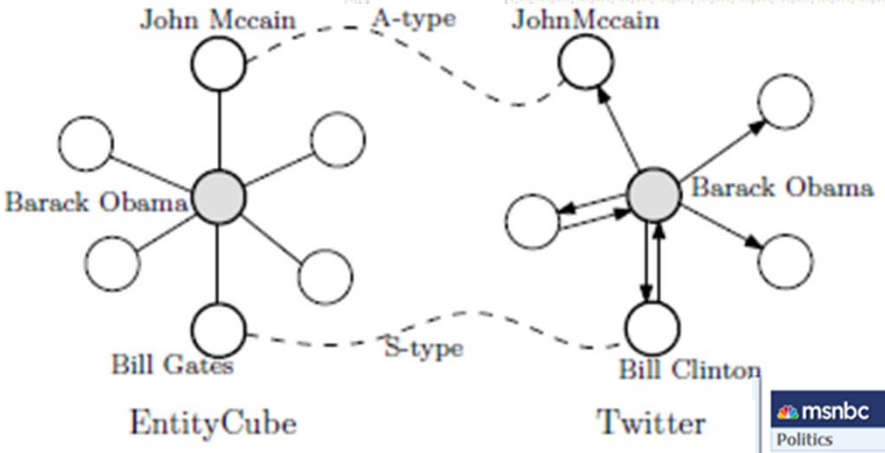
Name Steve Jobs
Location Apple
Bio More than meets the i. As you should expect from a parody account.

50 463,463 12,879

BarackObama

Verified Account

Name Barack Obama
Location Washington, DC
Web http://www.barack...
Bio 44th President of the United States



msnbc
Politics

Obama to Americans: 'We can be better'

NBC, msnbc.com and news services
7:23 PM EST January 13, 2011

Automatically generated link

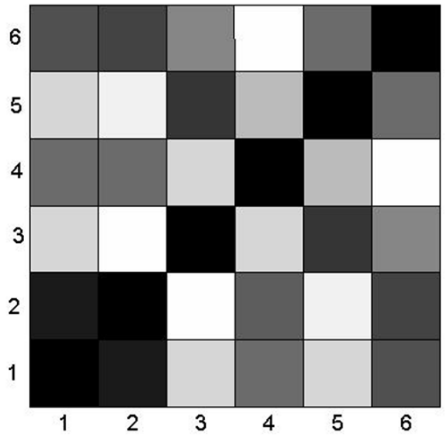
President Barack Obama mourned victims of the Arizona shooting spree on Wednesday and urged Americans not to let a political debate over tragedy be used as "one more occasion to turn on one another."

In a speech at a service for those killed in a weekend massacre that left Rep. Gabrielle Giffords gravely wounded, the president appealed for national unity and soul-searching after the shootings. He urged Americans to "expand our moral imaginations" and "sharpen our instincts fc empathy" - even with those who are political adversaries.

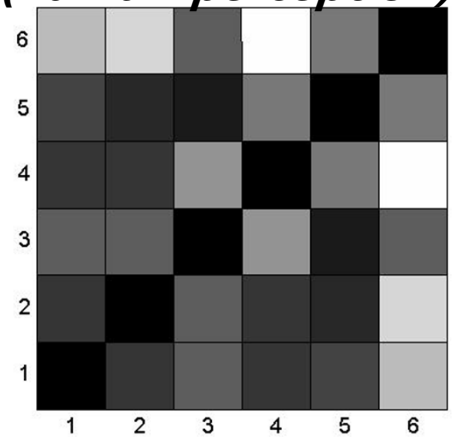
"I believe we can be better," Obama said to a capacity crowd at the University of Arizona basketball arena - and to countless others watchin America. "Those who died here, those who saved lives here - they help me believe. We may not be able to stop all evil in the world, but I k that how we treat one another is entirely up to us."

Six people were killed and 13 wounded in the massacre. Police allege 22-year-old Jared Lee Loughner shot Giffords and many in the line of

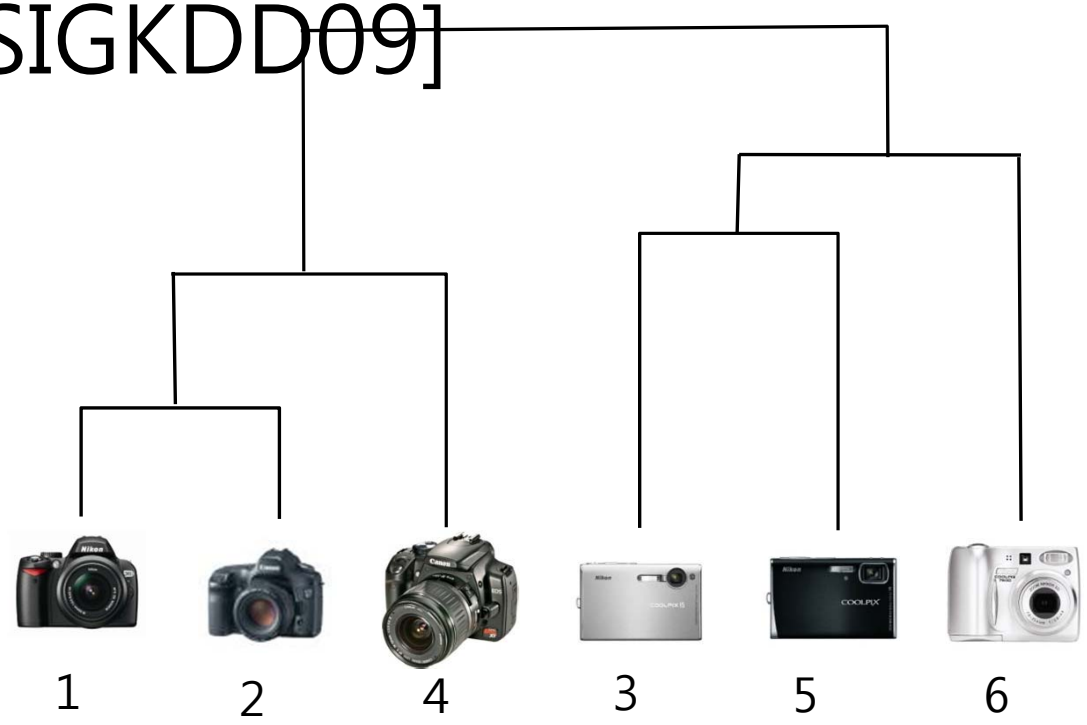
EntityCube and Feature Graph [SIGKDD09]



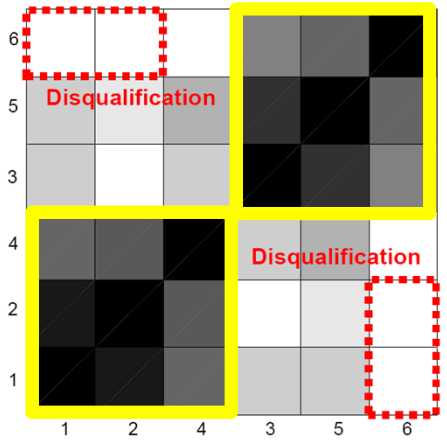
EntityCube (human perception)



Feature-based (value locality)

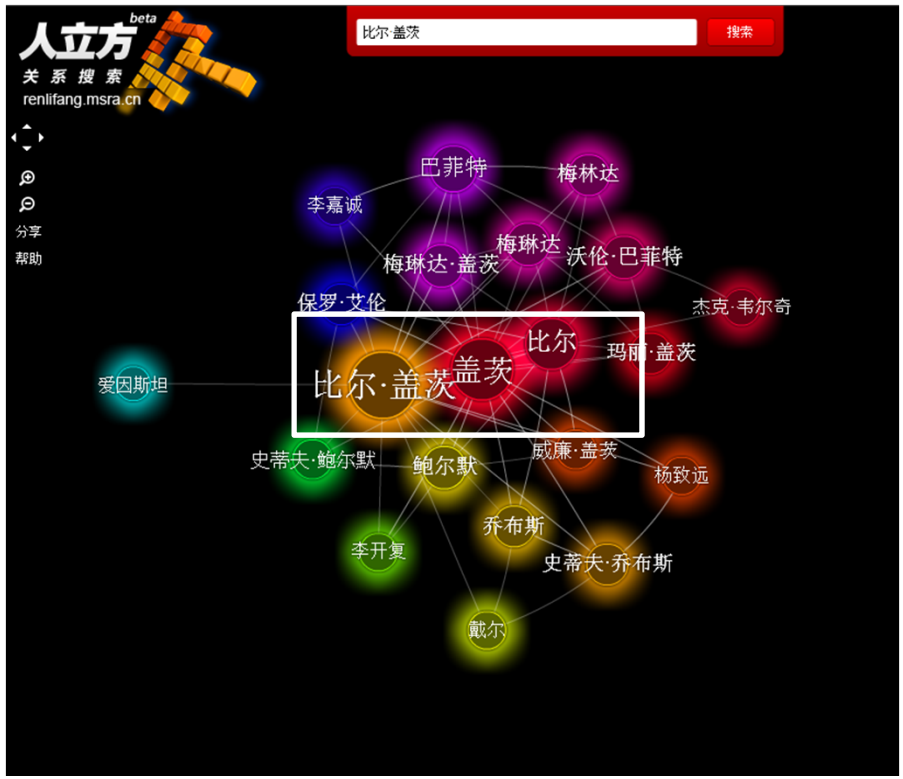


Clustering results



A Big Challenge Under the Rug

EntityCube



Probase

US Presidents such as Bill Clinton and George W. Bush

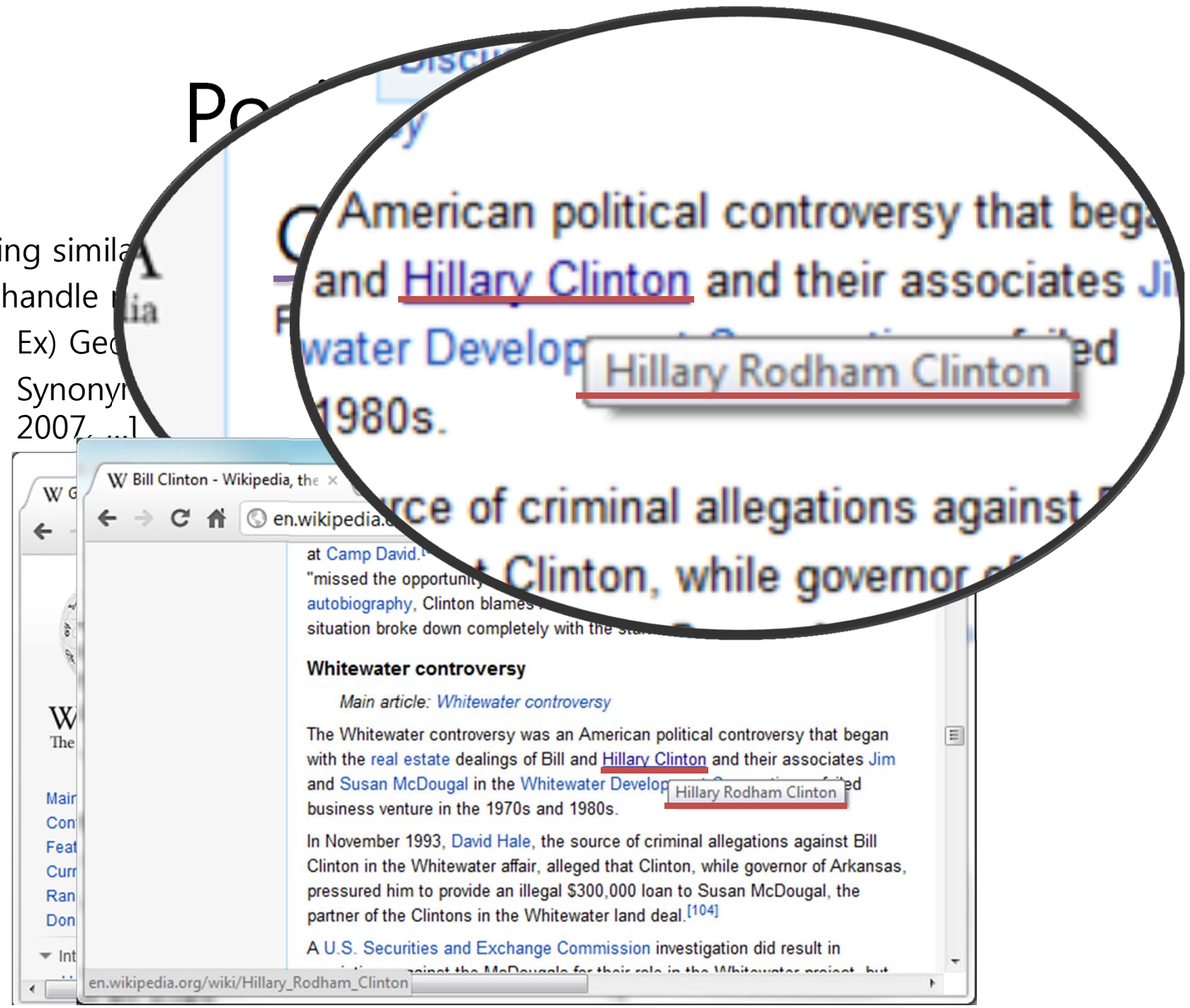
US Presidents
Bill Clinton
William J Clinton
George W. Bush
George H.W. Bush
Dubya

Web-scale Cleansing [VLDB11]

Joint work with Taesung Lee[1]
Zhongyuan Wang[2], Haixun Wang[2]

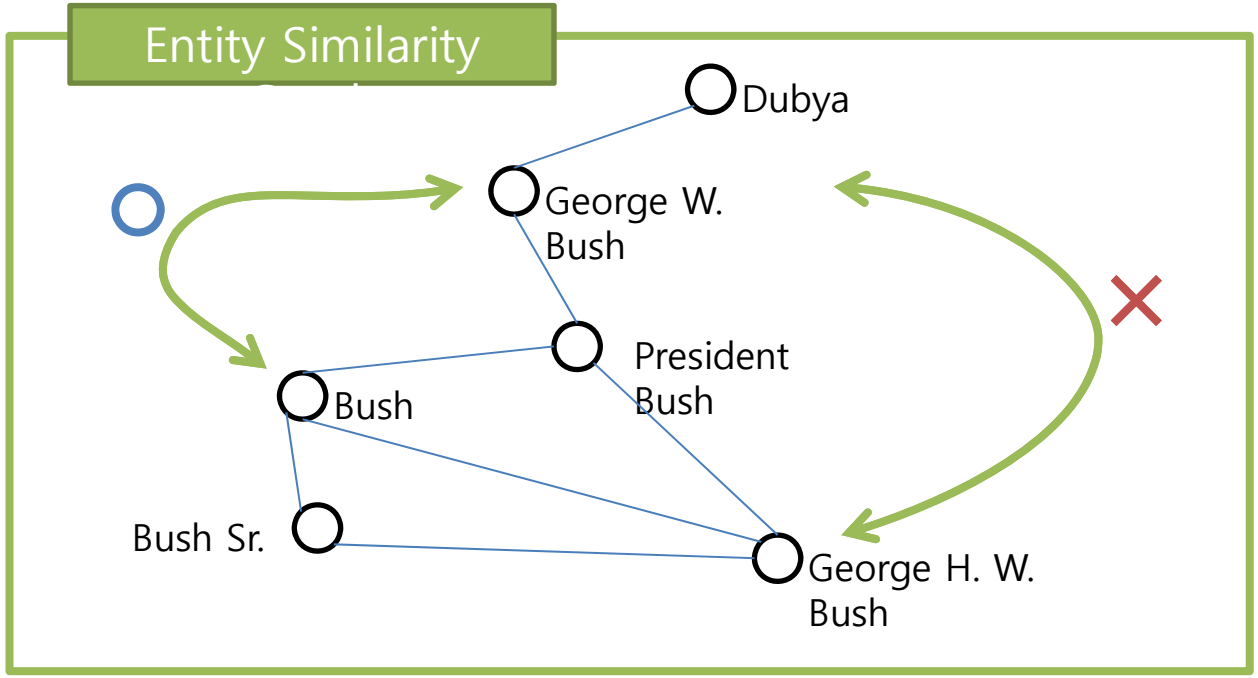
[1]Pohang University of Science and Technology (POSTECH), Korea
[2]Microsoft Research Asia (MSRA), China

- String simila
- To handle
  - Ex) Ge
  - Synony
    2007, ...]



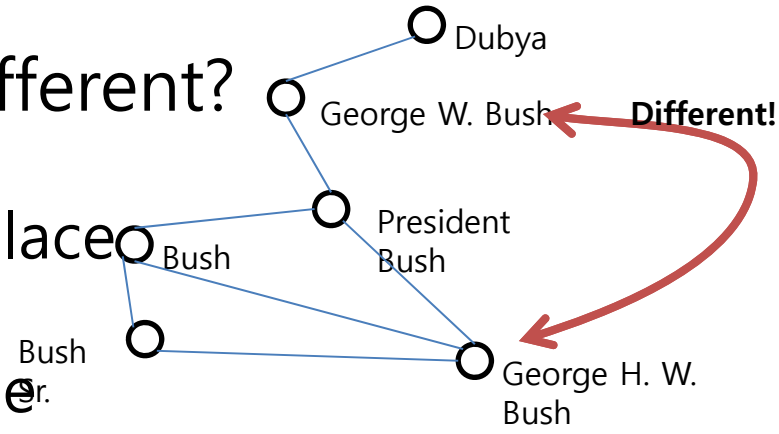
What If Just Positive Evidence?

- George W. Bush = President Bush
  + President Bush = George H. W. Bush
  = ?

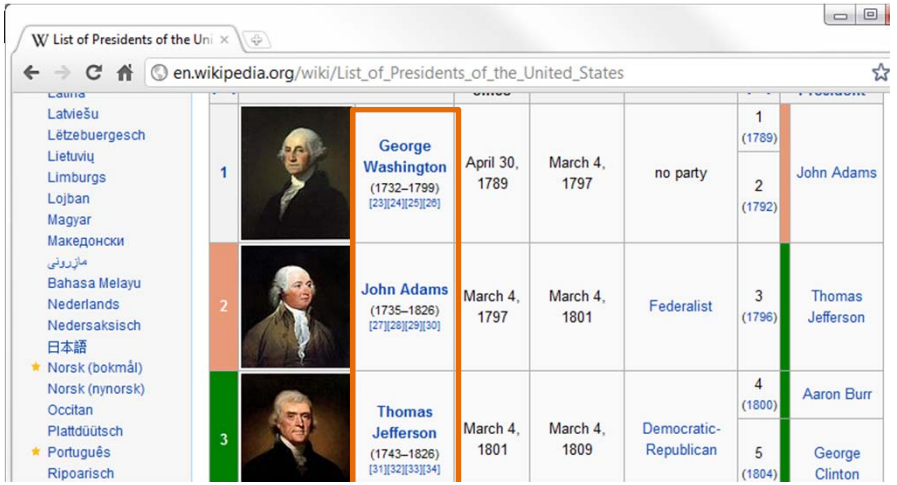


Negative Evidence

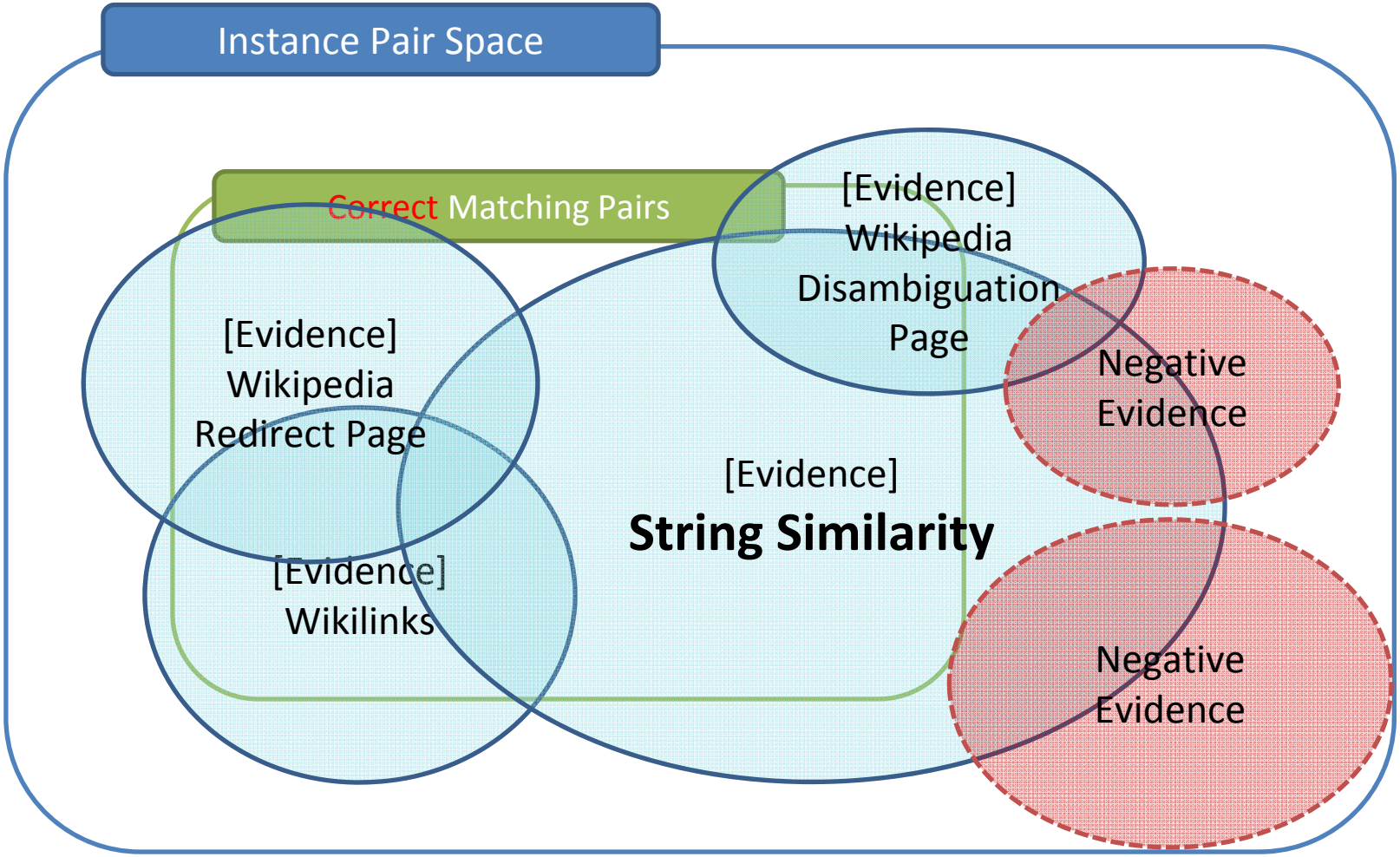
- What if we know they are different? It is possible to…
  - stop transitivity at the right place
  - safely utilize string similarity
  - remove false positive evidence



- *'Clean data has no duplicates'* Principle
  - The same entity would not be mentioned in a list of instances several times in different forms
  - We assume these are clea
    - 'List of ***', 'Table of ***'
    - such as …
    - Freebase

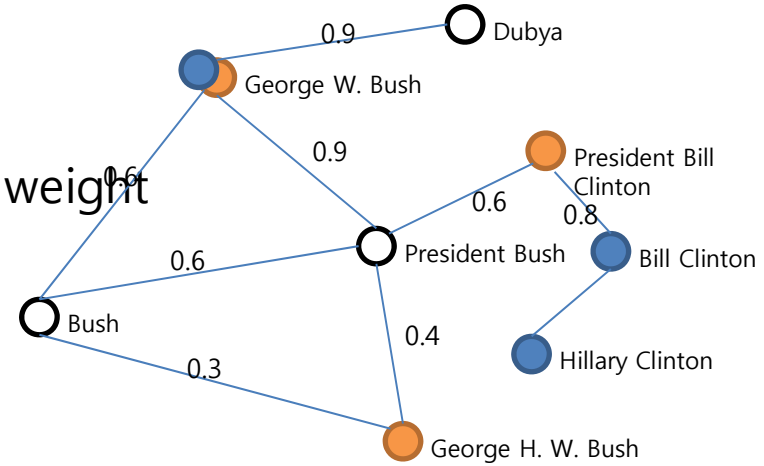


Two Types of Evidence in Action



Our Approach

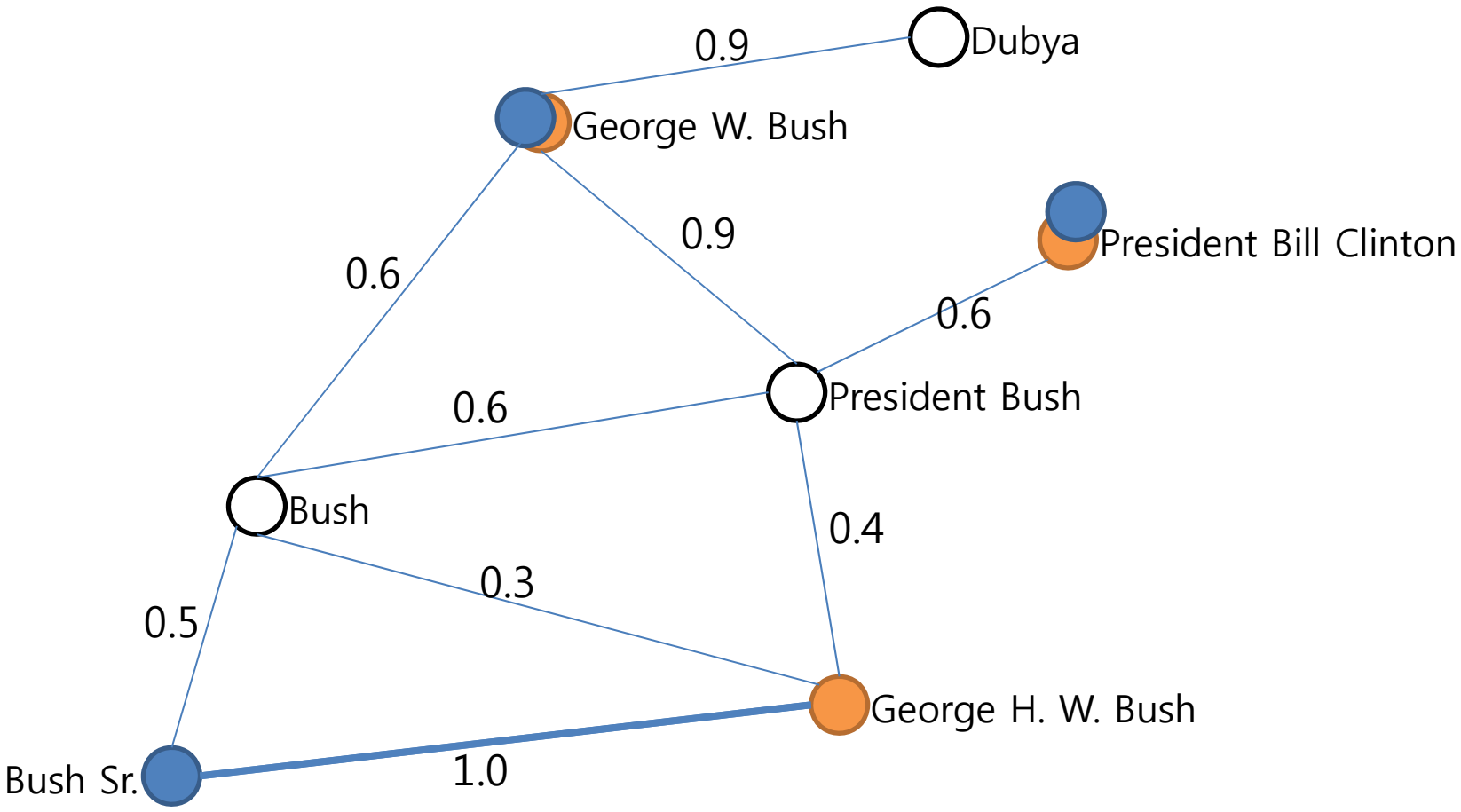
- Building a graph of entities
  - Quantifying positive evidence for edge weight
  - Noisy-or model
    - $1 - (1 - a_1) \cdot (1 - a_2) \cdots (1 - a_k)$
    - Using All kinds of positive evidence including string similarity

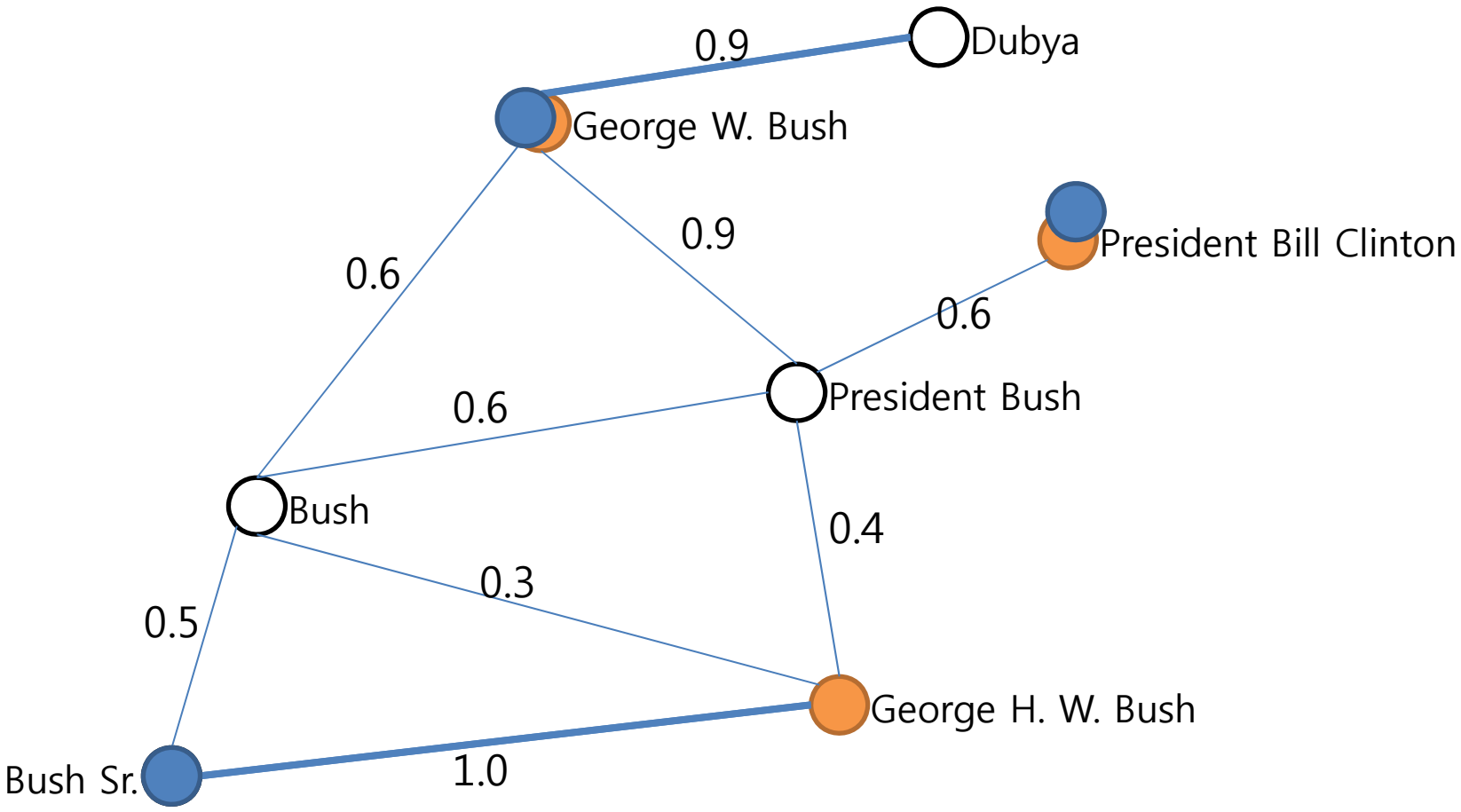


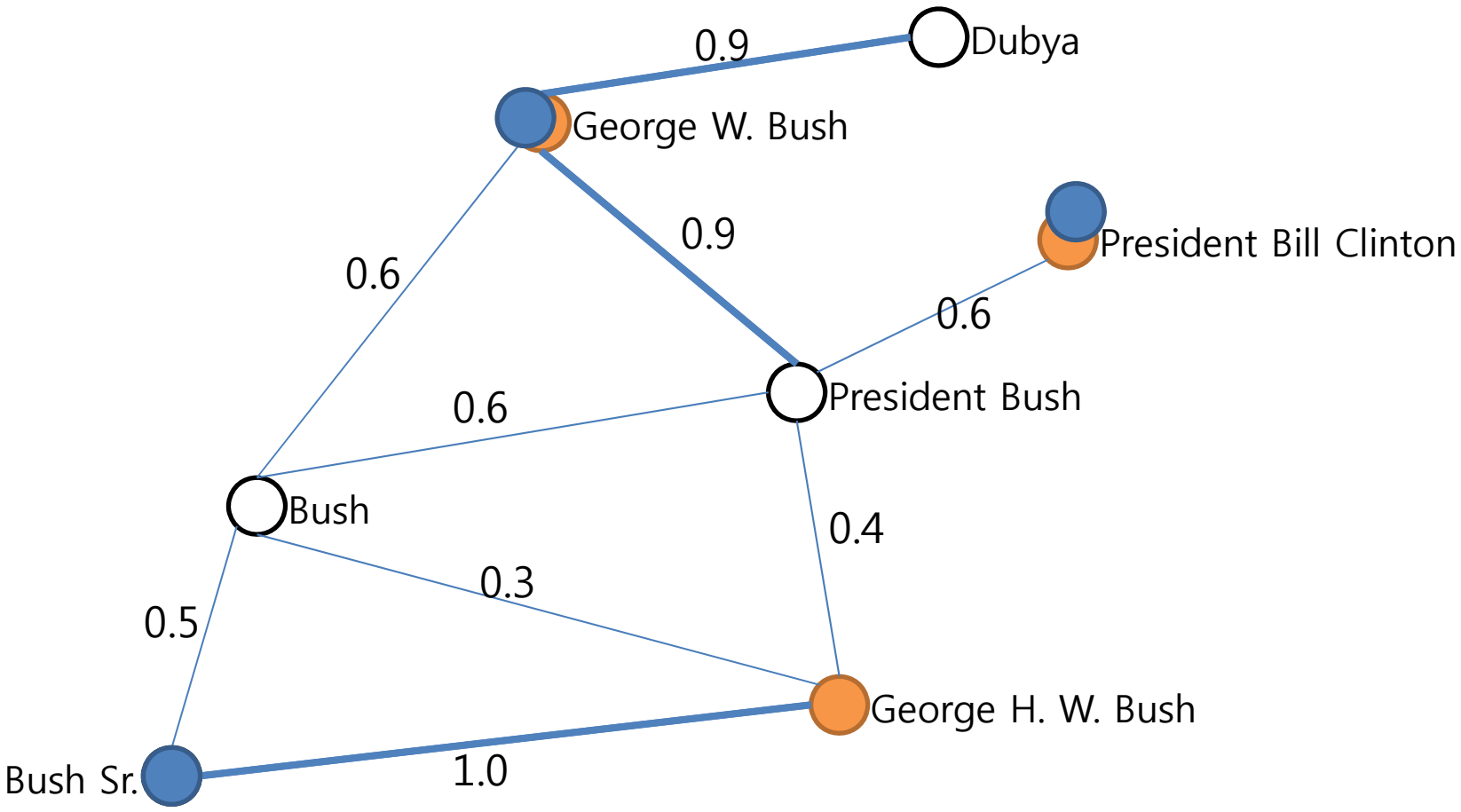
- Multi-Multiway Cut problem on a graph of entities
  - A variant of s-t cut
  - No vertices with the same label may belong to one connected component (cluster) with a cost function; the sum of removed edge weights
    - {George W. Bush, President Bill Clinton, George H. W. Bush}, {George W. Bush, Bill Clinton, Hillary Clinton}, …
  - Some instance may belong to any of several possible clusters
    - Bush to George W. Bush, or George H. W. Bush

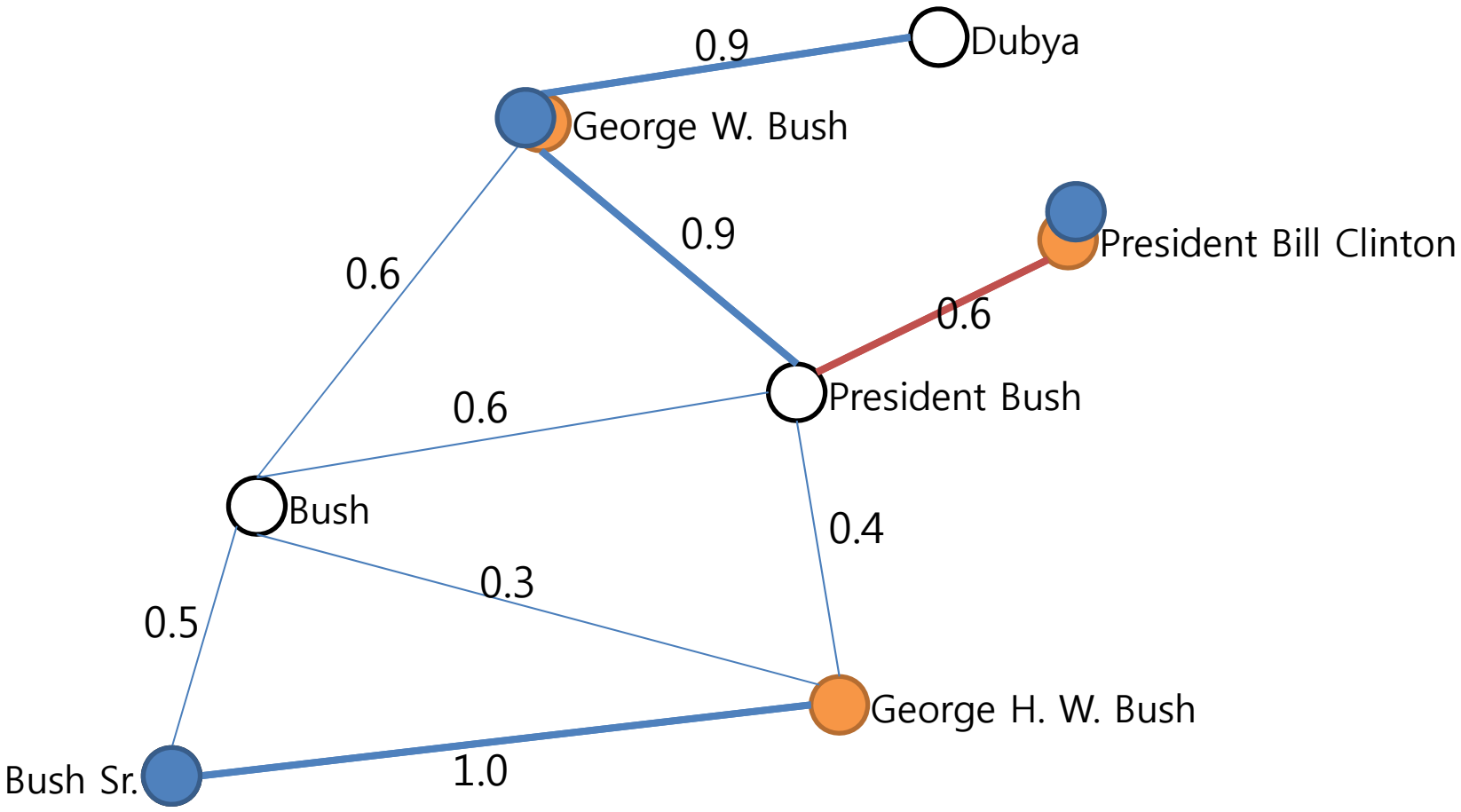
Our method

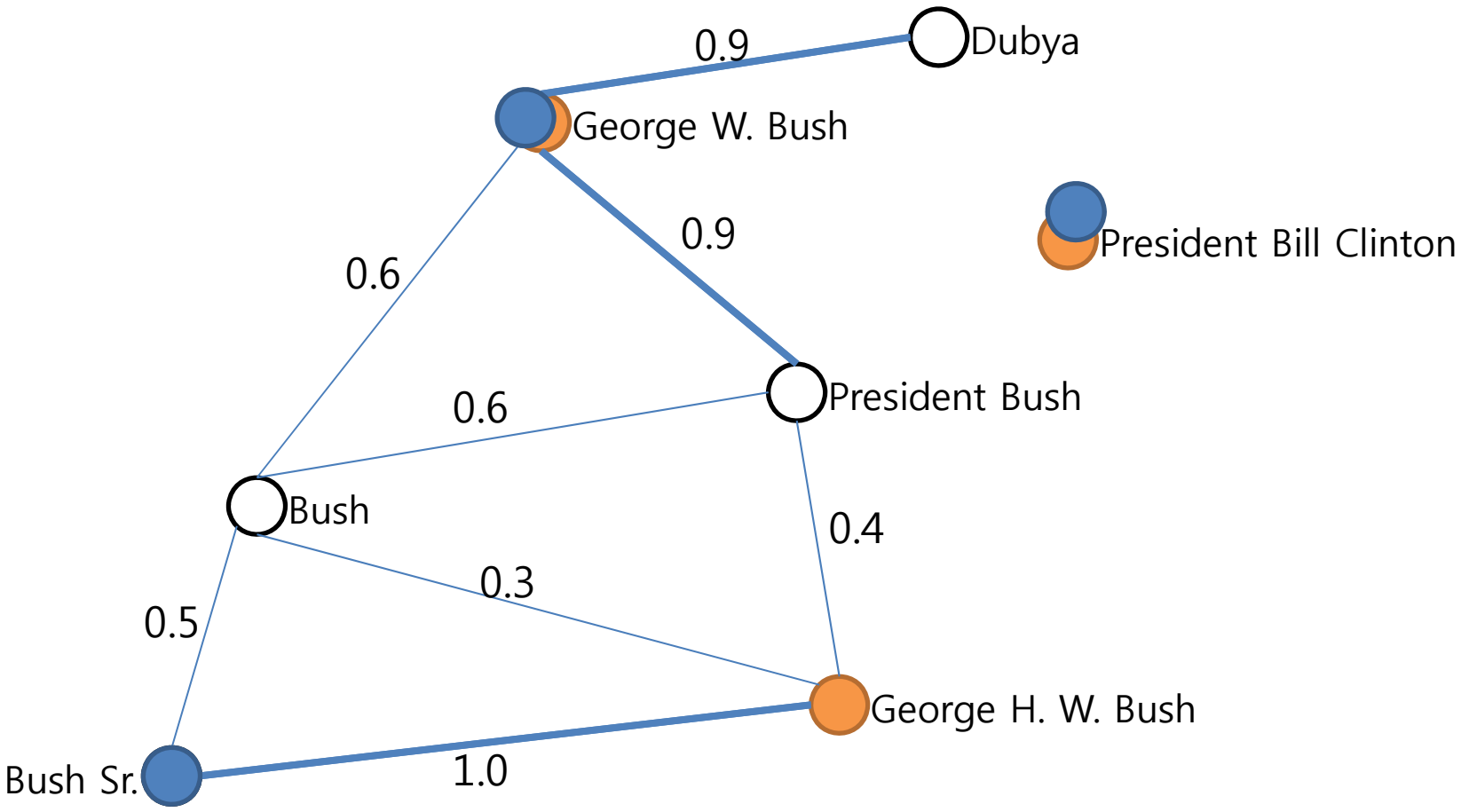
- Checking violation with each ‘monte-carlo’ insertion
  - Randomly insert one edge at a time with probability proportional to the weight of edge
  - Skip the edge that makes violation
- We repeat the random process for a fixed number of times, and then choose the best one that minimize the cost; sum of dropped edge weights

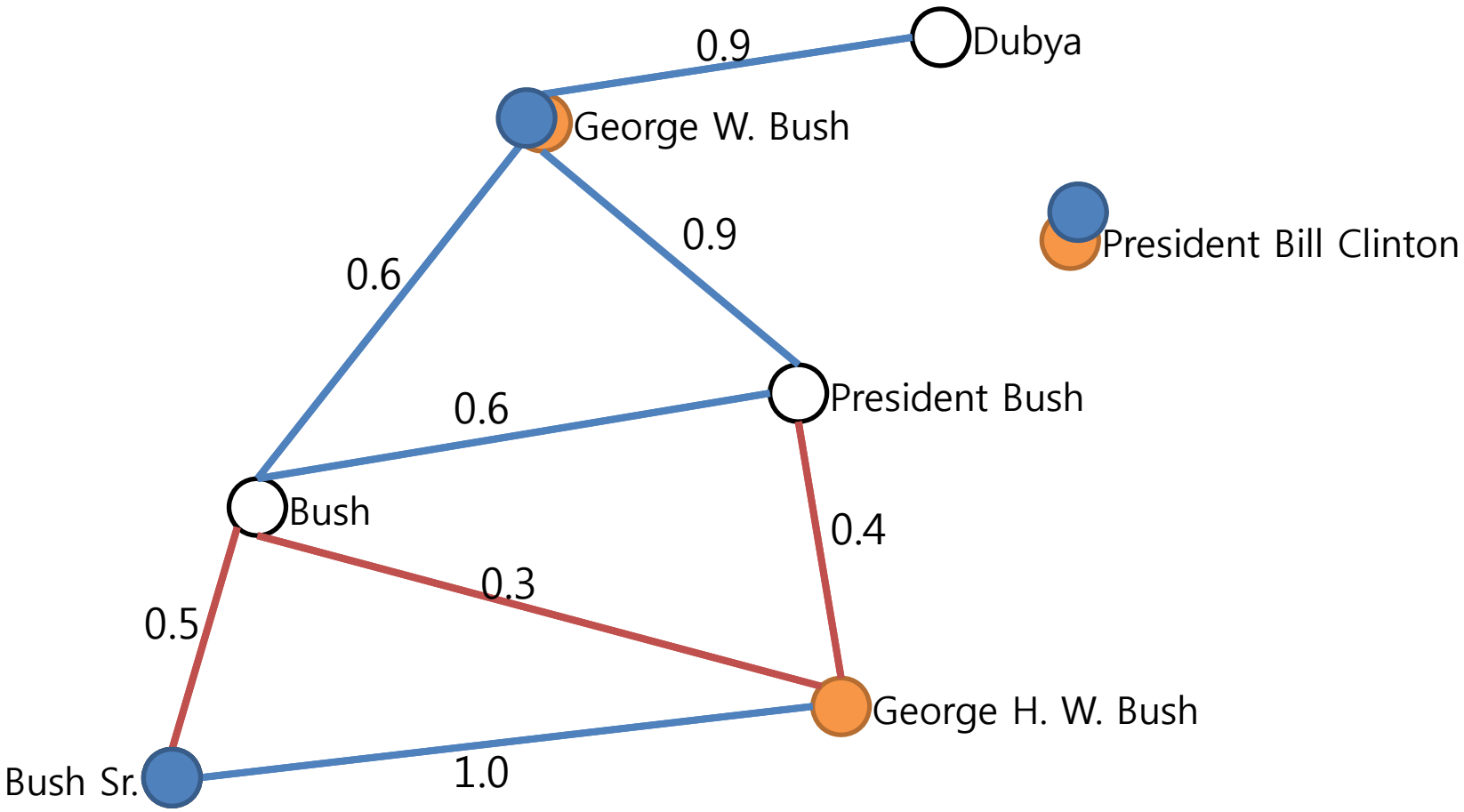

Dubya
0.9
George W. Bush
0.9
0.6
President Bill Clinton
0.6
0.6
President Bush
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0

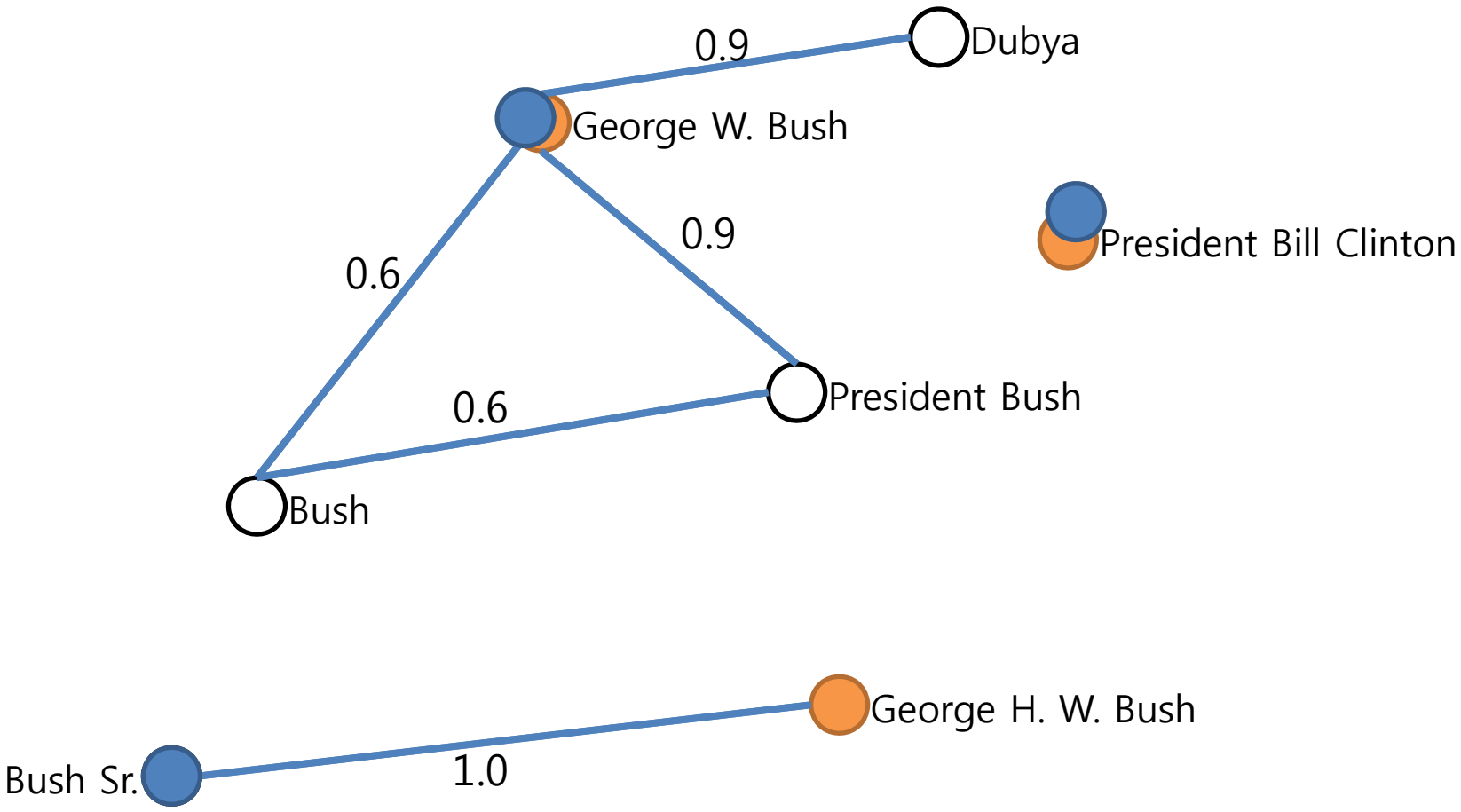

0.9
Dubya
George W. Bush
0.9
0.6
President Bill Clinton
0.6
0.6
President Bush
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0


0.9
Dubya
George W. Bush
0.9
0.6
President Bill Clinton
0.6
0.6
President Bush
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0


Dubya
0.9
George W. Bush
0.9
President Bill Clinton
0.6
0.6
President Bush
0.6
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0

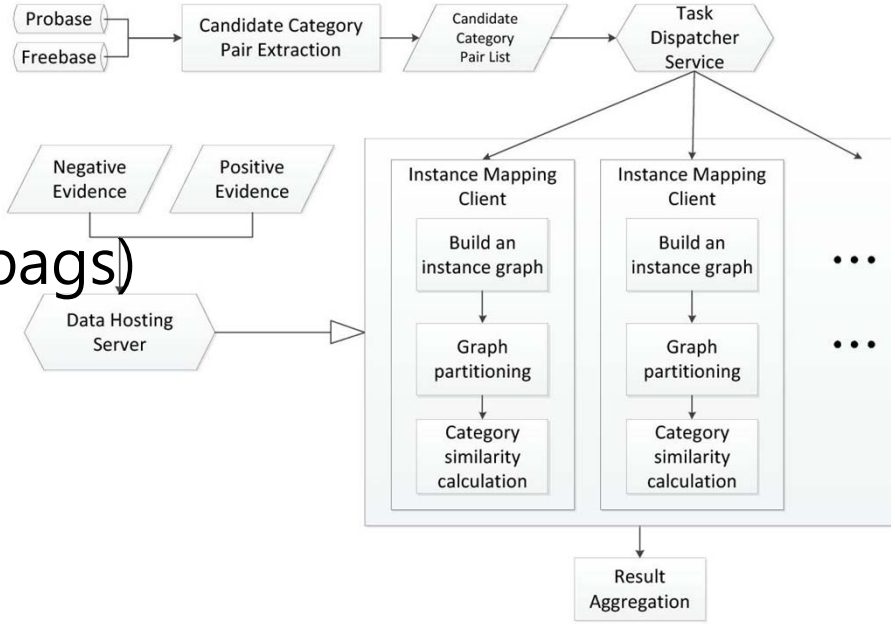

0.9
Dubya
George W. Bush
President Bill Clinton
0.9
0.6
0.6
President Bush
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0


0.9
Dubya
George W. Bush
President Bill Clinton
0.9
0.6
0.6
President Bush
Bush
0.4
0.3
0.5
George H. W. Bush
Bush Sr.
1.0


0.9
Dubya
George W. Bush
President Bill Clinton
0.9
0.6
0.6
President Bush
Bush
George H. W. Bush
Bush Sr.
1.0

Experimental Setting

- Probase: 16M instances / 2M concepts
- Freebase (2010-10-14 dump): 13M instances / 12.7K concepts
- Positive Evidence
  - String similarity: Weighted Jaccard Similarity
  - Wikipedia Links: 12,662,226
  - Wikipedia Redirect: 4,260,412
  - Disambiguation Pages: 223,171
- Negative Evidence (# name bags)
  - Wikipedia List: 122,615
  - Wikipedia Table: 102,731
  - Freebase



Experimental Setting

- Baseline #1
  - Positive evidence (without string similarity) based method.
  - Maps two instances with the strongest positive evidence. If there is no evidence, not mapped.
- Baseline #2
  - String similarity based method.
  - Maps two instances if the string similarity, here Jaro-Winkler distance, is above a threshold (0.9 to get comparable precision).
- Baseline #3
  - Markov Clustering with our similarity graphs
  - The best method among the scalable clustering methods for entity resolution introduced in [Hassanzadeh, VLDB2009]

Some Examples of Mapped Probase Instances

Freebase Instance	Baseline #1	Baseline #2	Baseline #3	Our Method
George W. Bush	George W. Bush, President George W. Bush, Dubya	George W. Bush	George W. Bush, President George W. Bush, George H. W. Bush, George Bush Sr., Dubya	George W. Bush, President George W. Bush, Dubya, **Former President George W. Bush**
American Airlines	American Airlines, American Airline, AA, Hawaiian Airlines	American Airlines, American Airline	American Airlines, American Airline, AA, North American Airlines	American Airlines, American Airline, AA
John Kerry	John Kerry, Sen. John Kerry, Senator Kerry	John Kerry	John Kerry, Senator Kerry	John Kerry, Sen. John Kerry, Senator Kerry, **Massachusetts Sen. John Kerry, Sen. John Kerry of Massachusetts**

Some Examples of Mapped Probase Instances

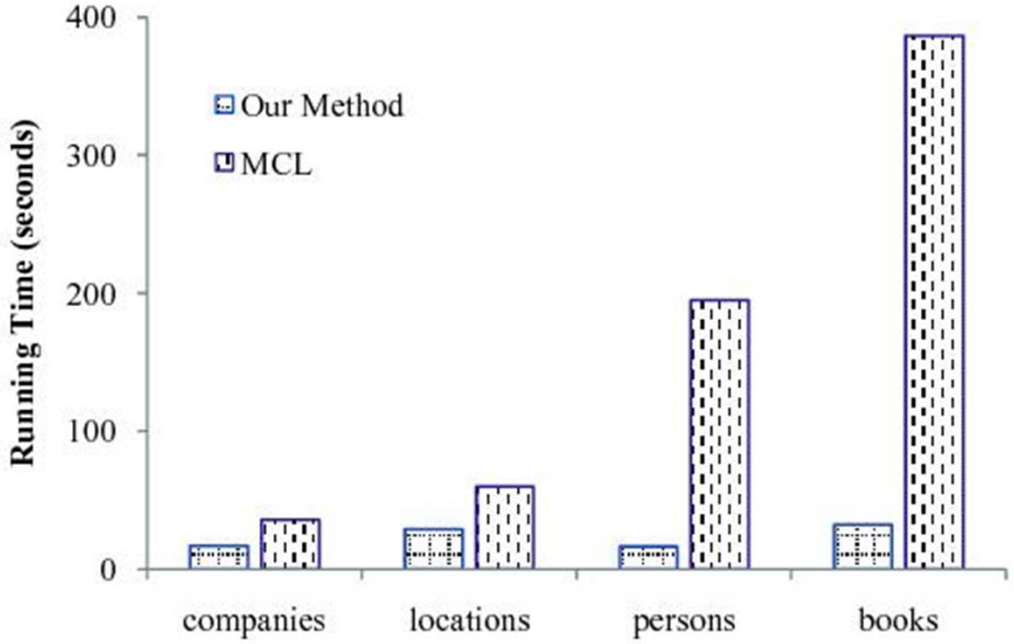
- Baseline #1 cannot deal with misspelled instances and variations → low recall
  - Massachusetts Sen. John Kerry
  - Sen. John Kerry of Massachusetts
- Baseline #2 produces false positives & shows weakness on large length variations
  - George W. Bush ~ George H. W. Bush
  - Hillary Clinton ~ Bill Clinton
- Baseline #3 produces false positives by wrong transitivity
  - George W. Bush ~ George H. W. Bush
  - American Airlines ~ North American Airlines

Precision / Recall

Probase Class	**Politicians**		**Foramts**		**Systems**		**Airline**	
Freebase Type	**/government /politician**		**/computer /file_format**		**/computer /operating_system**		**/aviation /airline**	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Baseline #1	**0.99**	0.66	0.90	0.25	0.90	0.37	0.92	0.63
Baseline #2	0.97	0.31	0.79	0.27	0.59	0.26	0.96	0.47
Baseline #3	0.88	0.63	0.82	0.48	0.88	0.52	0.96	0.54
Our method	0.98	**0.84**	**0.93**	**0.86**	**0.91**	**0.59**	**1.00**	**0.70**

Scalability

- Our method (in C#) is compared with Baseline #3 (Markov Clustering, in C)



Category	|V|	|E|
Companies	662,029	110,313
Locations	964,665	118,405
Persons	1,847,907	81,464
Books	2,443,573	205,513

Thanks

- http://www.postech.ac.kr/~swhwang