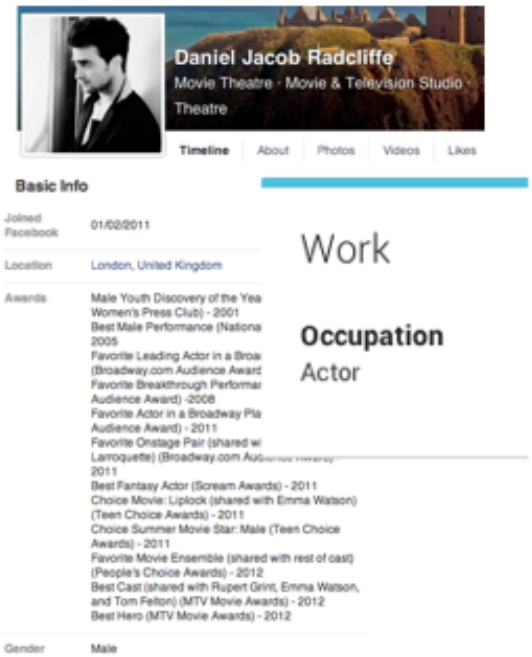
Weakly Supervised User Profile Extraction from Twitter

Jiwei Li, Alan Ritter and Eduard Hovy

School of Computer Science, Carnegie Mellon University

June 22nd, 2014

User Profile



Education

Job

Interest

Gender

Spouse/Lover

Favorite Movie

Religion

Favorite Food

Main Contribution

Automatic extraction of attributes from Twitter:

- Spouse
- Education
- Job

Main Contribution

- We use Google plus as distant supervision for user attribute extraction.
- We present a large-scale dataset for this task.
- We demonstrate the benefit of jointly reasoning about users' social network structure.

Outline

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

Motivation/Introduction

Motivation/Introduction

Work and Education

Stanford University
Professor · Stanford, California · Sep 2009 to present

Cornell University
Post Doc · Ithaca, New York · Sep 2008 to Sep 2009

See All Employers

Cornell University
Ithaca, New York

Carnegie Mellon University
PhD in Machine Learning · Machine Learning, AI · Computer Science · Pittsburgh, Pennsylvania

University of Ljubljana
faculty of computer and information science · Ljubljana, Slovenia

Gimnazija Bezigrad
Class of 1999 · Ljubljana, Slovenia

Places Lived

Palo Alto, California
Current City

Sentjost nad Horjulom
Hometown

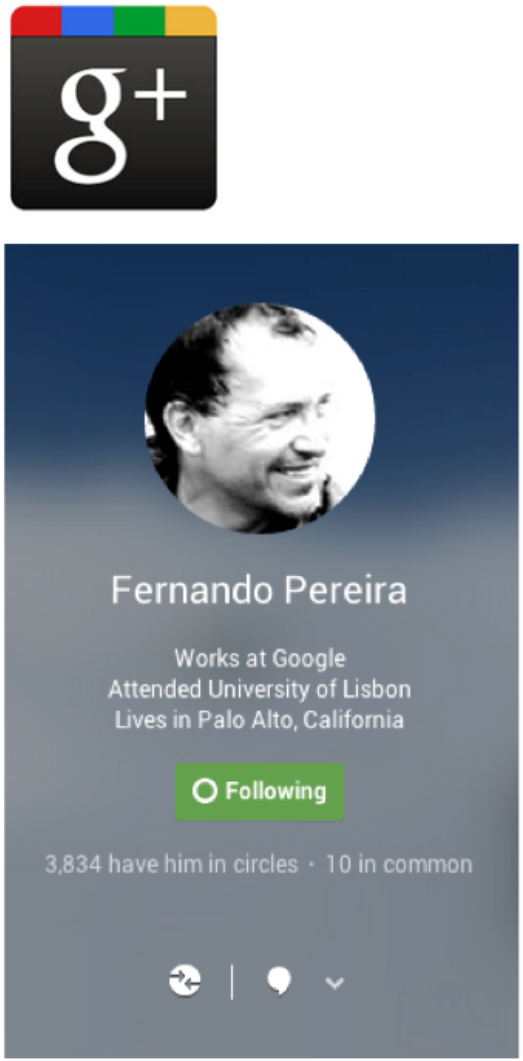
Basic Information

Religious Views	Catholic

Contact Information

Mobile Phones	(650) 725-3711
Address	353 Serra Mall Stanford, CA, United States 94305
Website	http://cs.stanford.edu/~jure
Facebook	http://facebook.com/jure.leskovec

Motivation/Introduction



Work

Occupation
Distinguished researcher

Employment

Google
Research director, 2008 - present

LNEC, Lisbon
1975 - 1977

SRI International
1982 - 1989

AT&T
1989 - 2000

University of Pennsylvania
2001 - 2007

Education

University of Lisbon
Mathematics, 1969 - 1975

University of Edinburgh
Artificial intelligence, 1977 - 1982

Basic Information

Gender Male

Relationship Married

Motivation/Introduction

Cathy Meng Xue

Investment Banking Analyst at Deutsche Bank

New York, New York (Greater New York City Area) | Investment Banking

Join LinkedIn and access Cathy Meng Xue's full profile. It's free!

As a LinkedIn member, you'll join 300 million other professionals who are sharing connections, ideas, and opportunities.

- See who you and **Cathy Meng Xue** know in common
- Get introduced to **Cathy Meng Xue**
- Contact **Cathy Meng Xue** directly

View Cathy Meng's full profile

Cathy Meng Xue's Education

MIT Sloan School of Management
Master of Finance
2012 – 2013
•Recipient of Dean's MFin Fellowship in recognition of professional, academic and personal achievements

Cornell University
Bachelor of Arts (B.A.), Economics
2009 – 2012
Grade: 3.8/4.0

Cathy Meng Xue's Experience

Investment Banking Analyst
Deutsche Bank
Public Company; 10,001+ employees; DB; Investment Banking industry
July 2013 – Present (1 year) | Greater New York City Area
Natural Resources Group - Power & Utilities

Investment Banking Summer Analyst
J.P. Morgan
Public Company; 10,001+ employees; JPM; Financial Services industry
June 2011 – August 2011 (3 months) | Hong Kong

Marketing Intern
KGRA Energy, LLC
Privately Held; 1-10 employees; Renewables & Environment industry
January 2011 – January 2011 (1 month) | Greater Chicago Area

Cathy Meng Xue's Skills & Expertise

Bloomberg | Financial Analysis | Microsoft Excel | DCF Valuation | Financial Modeling | Investments

Motivation/Introduction

Why Profile Extraction ?

Motivation/Introduction

Why Profile Extraction ?

- Friend Recommendation

Motivation/Introduction

Why Profile Extraction ?

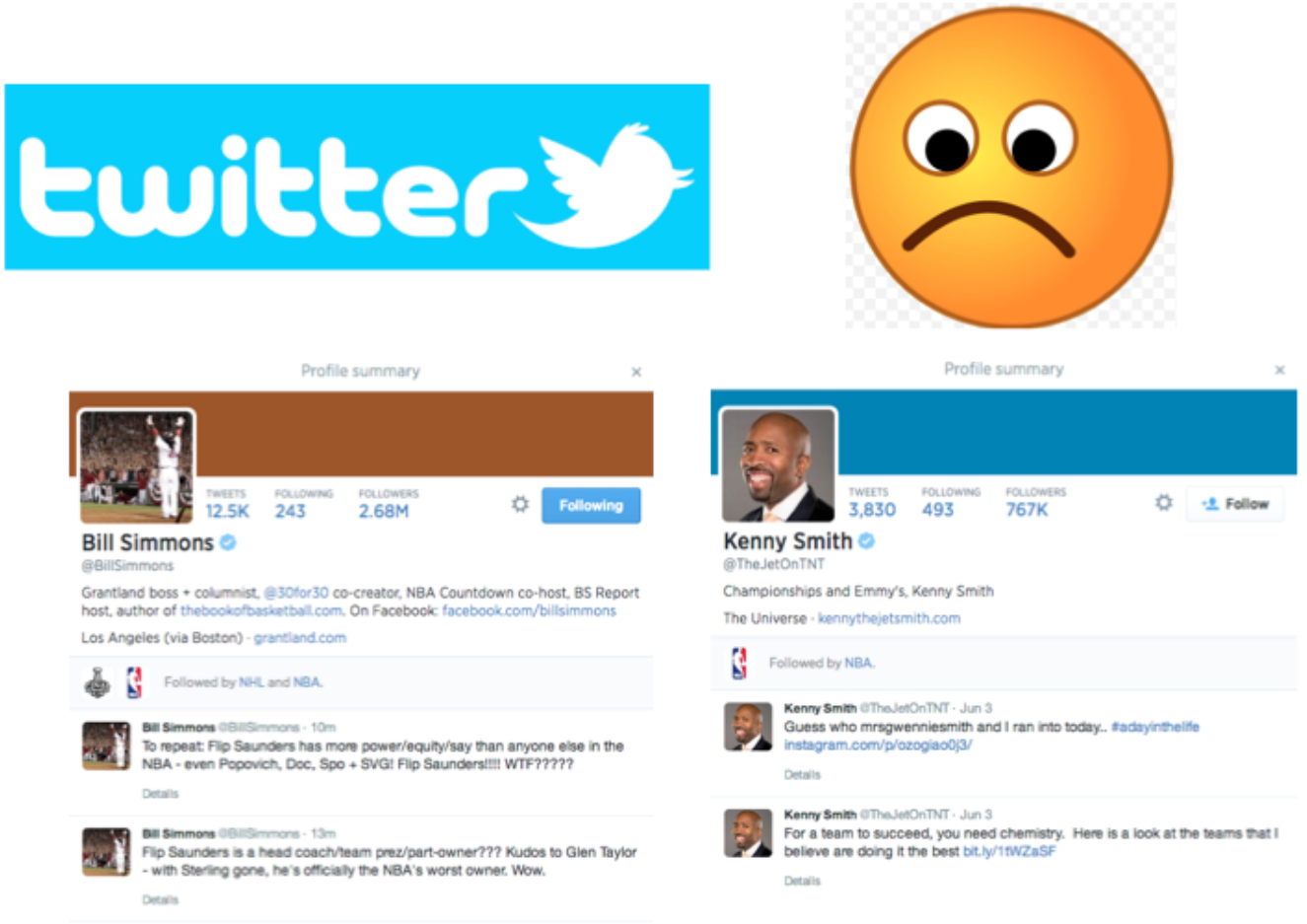
- Friend Recommendation
- Target Advertising (Movie, Book ...)

Motivation/Introduction

Why Profile Extraction ?

- Friend Recommendation
- Target Advertising (Movie, Book ...)
-

Motivation/Introduction



Motivation/Introduction

Twitter serves as a wonderful source:

Motivation/Introduction

Twitter serves as a wonderful source:

- Text Level Evidence

Motivation/Introduction

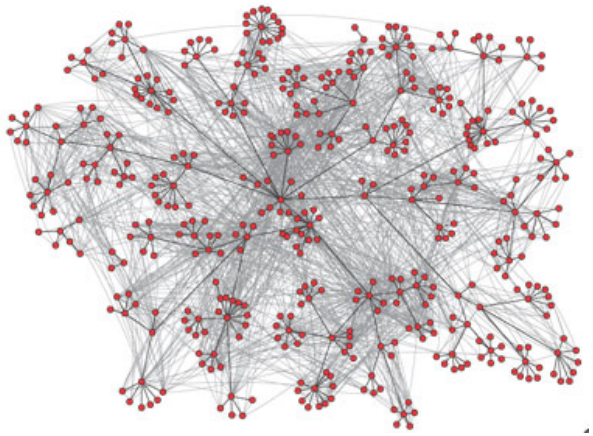
Twitter serves as a wonderful source:

- Text Level Evidence
- Network Information

Motivation/Introduction

Twitter serves as a wonderful source:

- Text Level Evidence
- Network Information
  - Homophily: People sharing more attributes have a higher chance of becoming friends in social media



Motivation/Introduction

Question

Unstructured Twitter data $\rightarrow$ Structured User Profile ?

Related Work

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

Related Work

User Attribute Extraction/ Identification

Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)

Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
- Age (Rao et al., 2010)

Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
- Age (Rao et al., 2010)
  - Relying on Amazon Mechanical Turk

Related Work

User Attribute Extraction/ Identification

- Gender (Ciot et al., 2013; Liu and Ruths, 2013)
- Age (Rao et al., 2010)
  - Relying on Amazon Mechanical Turk
- Political Polarity (Pennacchiotti et al, 2011)
  - Relying on external political websites

Dataset Creation

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

Dataset Creation

Challenge:

- Lack of Training Data

Dataset Creation

Distant Supervision

Dataset Creation

Distant Supervision

Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

IsCapital (Paris, France)
IsCapital (London, Britain)

Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

  - **Paris** is the capital and most populous city of **France**
  - The capital of **France** is **Paris**

Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

  - **Paris** is the capital and most populous city of **France**
  - The capital of **France** is **Paris**

Dataset Creation

What is Knowledge Base for our task ?

Dataset Creation

What is Knowledge Base for our task ?

Dataset Creation

What is Knowledge Base for our task ?

Dataset Creation

Attributes we focus on:

- Education
- Job
- Spouse

Dataset Creation

Attributes we focus on:

- Education
- Job
- Spouse

Dataset Creation

- Education: Positive Examples

Google Plus API

Dataset Creation

- Education: Positive Examples

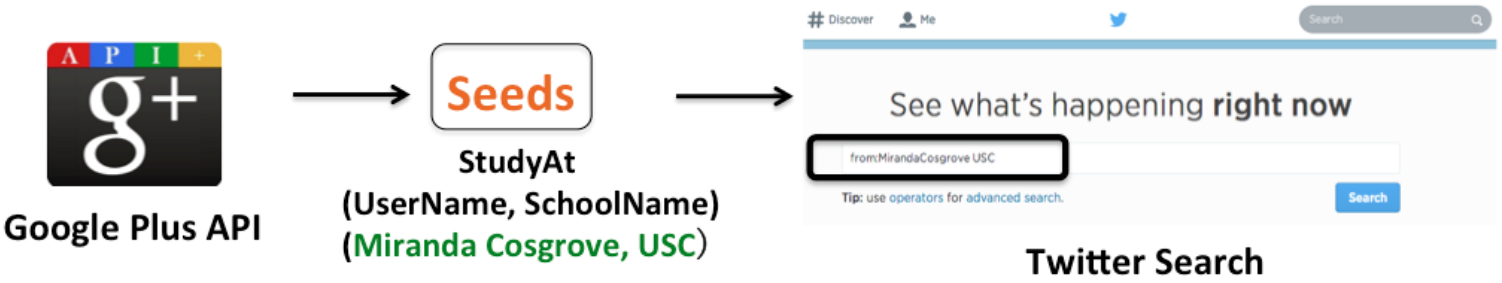
Google Plus API

Seeds

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

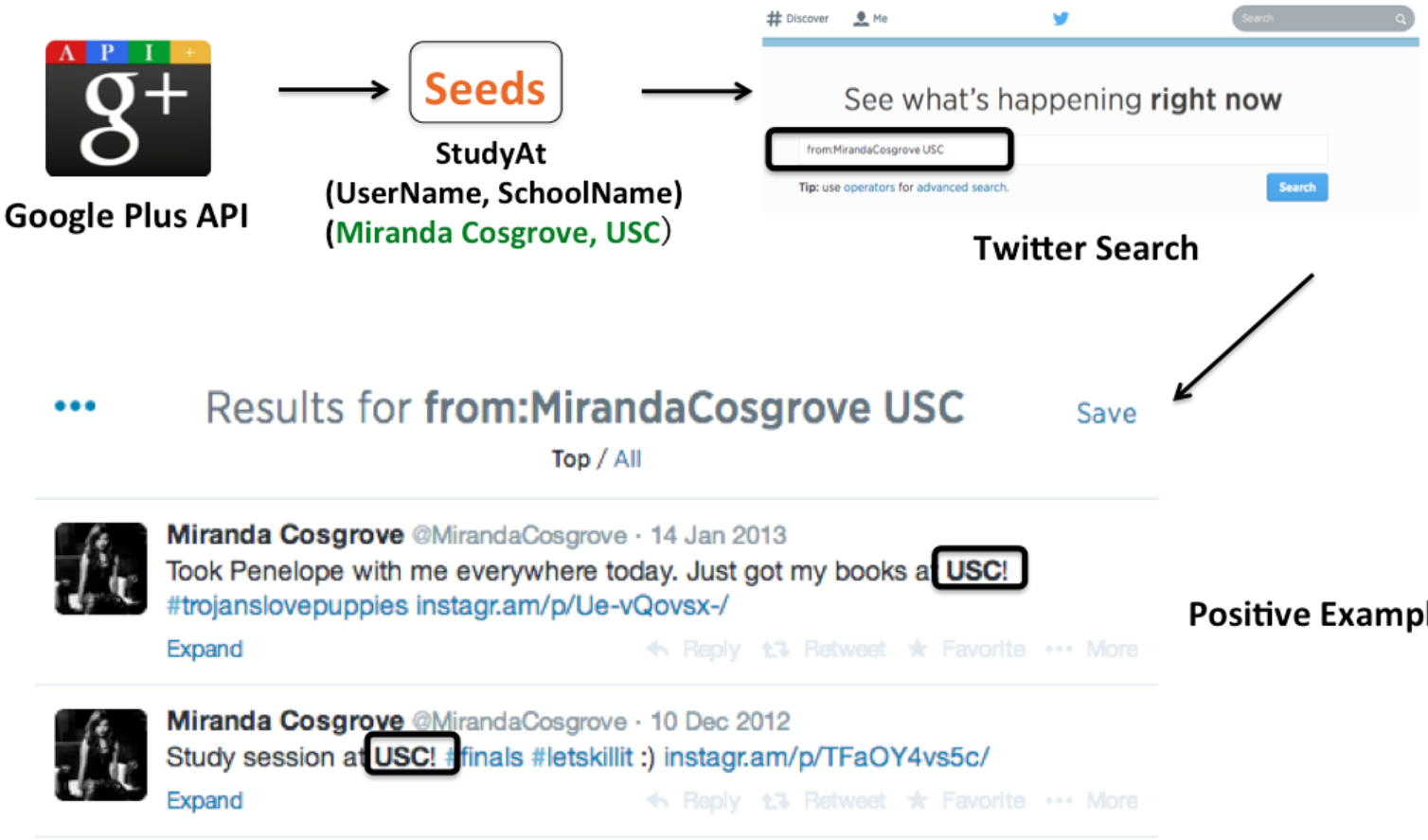
Dataset Creation

- Education: Positive Examples



Dataset Creation

- Education: Positive Examples



Dataset Creation

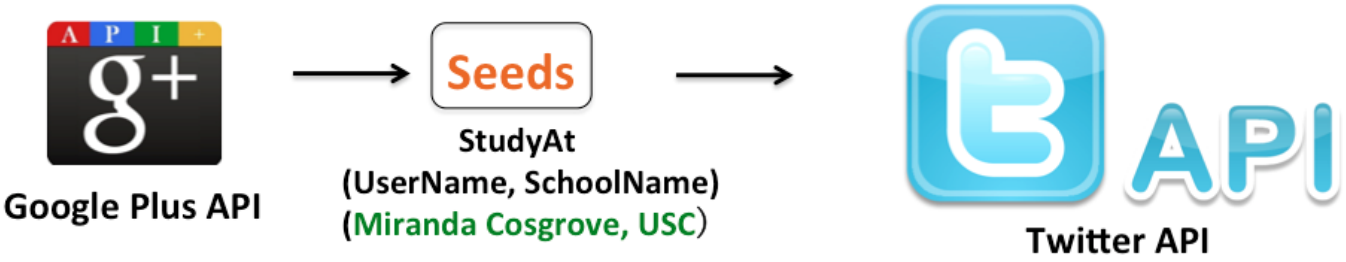
- Education: Negative Examples

Google Plus API

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

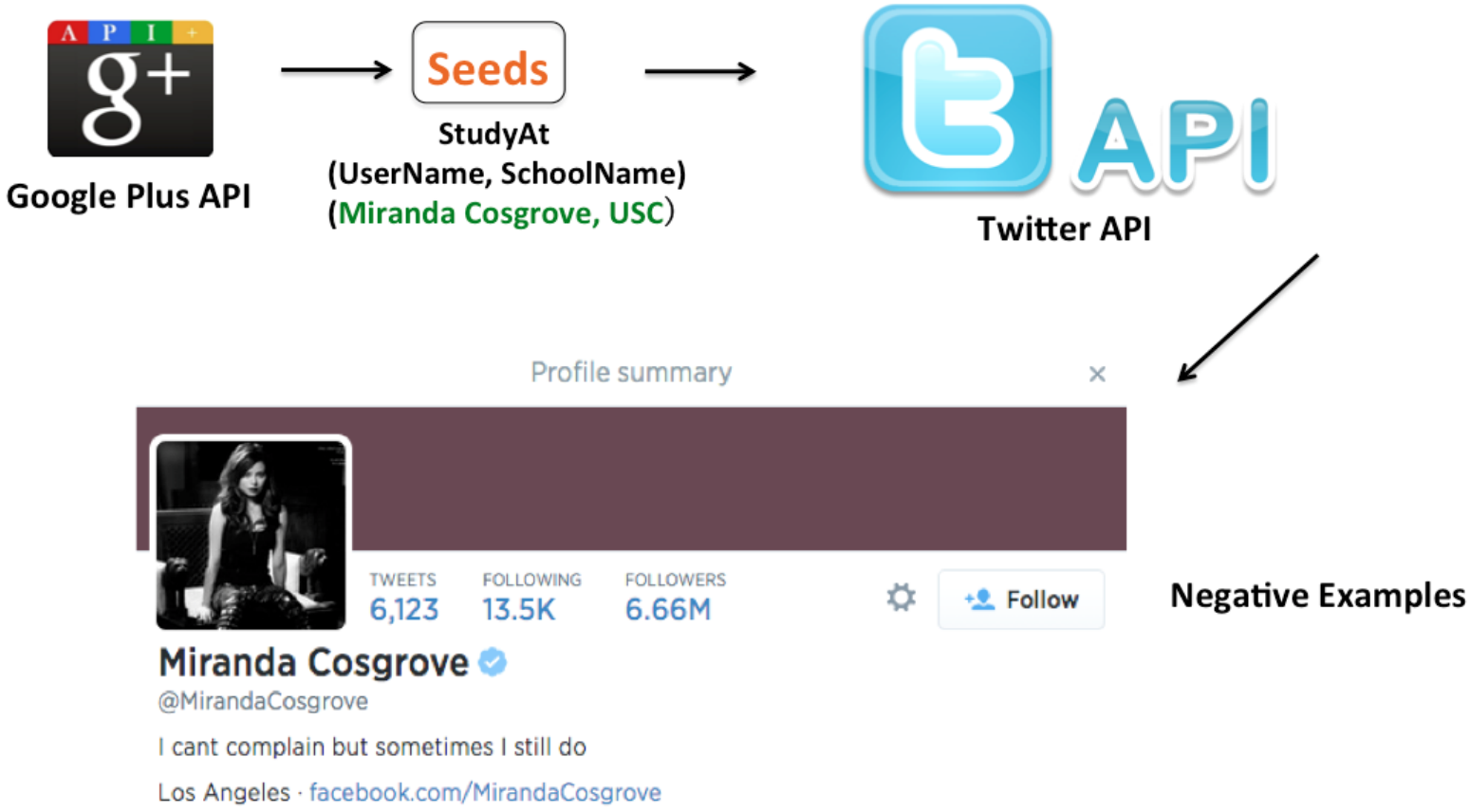
Dataset Creation

- Education: Negative Examples



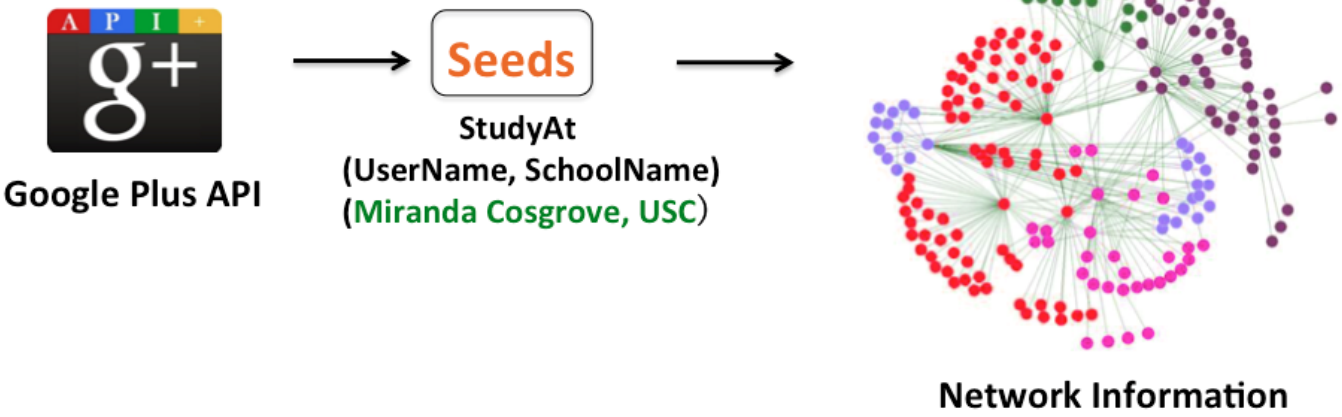
Dataset Creation

- Education: Negative Examples



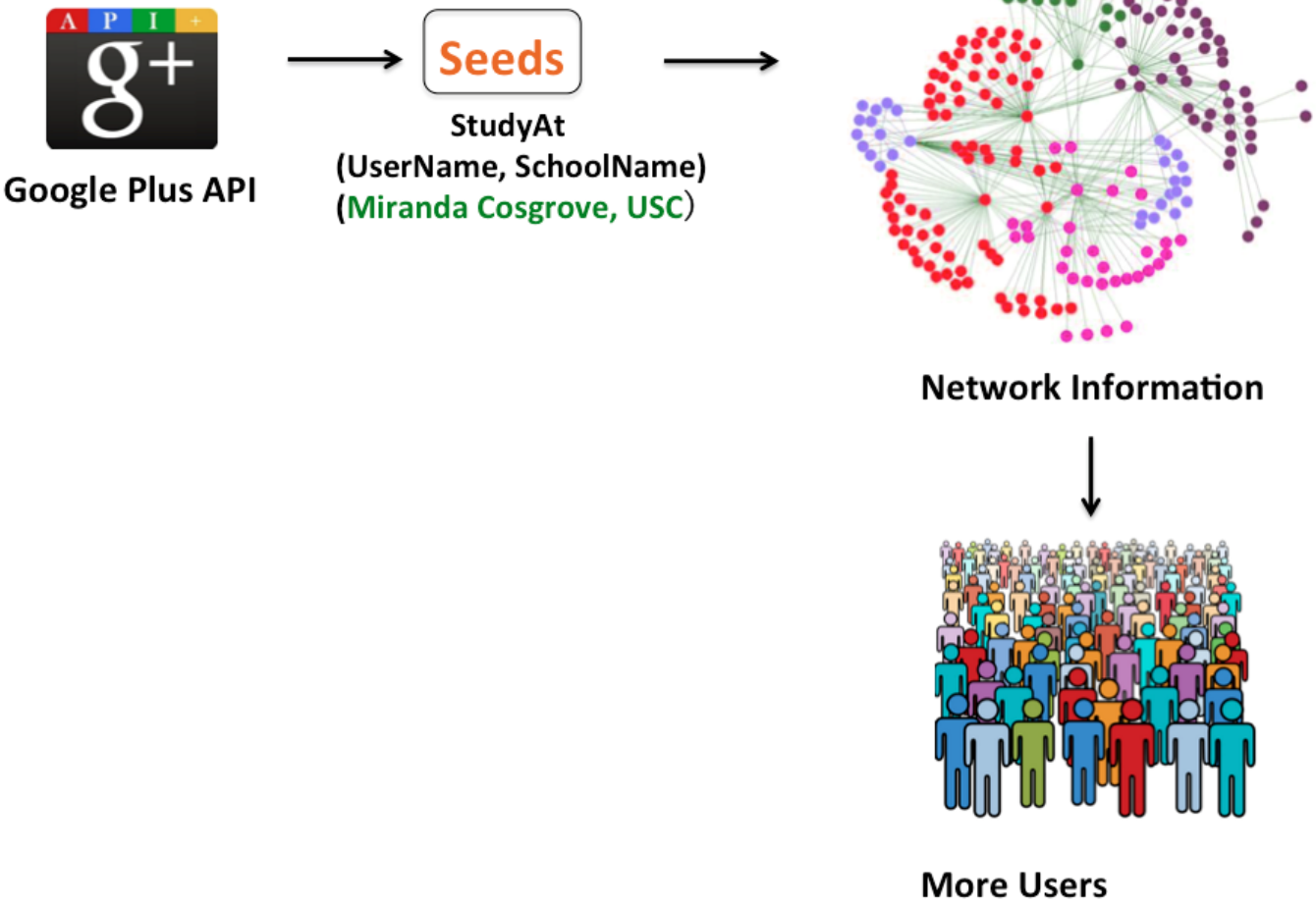
Dataset Creation

- Education: Data Expansion



Dataset Creation

- Education: Data Expansion



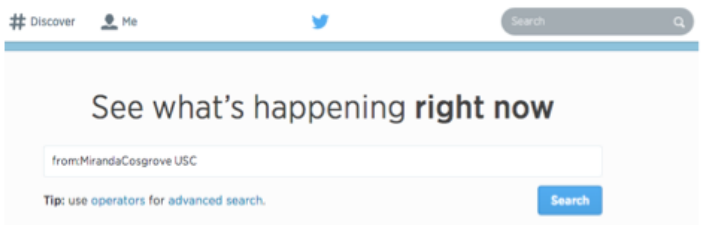
Dataset Creation

- Education: Data Expansion

Google Plus API

StudyAt
(UserName, SchoolName)
(Miranda Cosgrove, USC)

Network Information



More Users

Dataset Creation

- Spouse

Freebase API

Dataset Creation

- **Spouse**

Freebase API

IsSpouse
(Tom Cruise, Katie Holmes)


Dataset Creation

- Spouse

Freebase → Freebase API

IsSpouse
(Tom Cruise, Katie Holmes)


Tom Cruise @TomCruise · 19 Dec 2011
Excitement on the red carpet as **Katie Holmes** walks by at the Ghost Protocol premiere!! #MissionImpossible yfrog.us/5abj8z
Expand Reply Retweet Favorite More

Tom Cruise @TomCruise · 1 Oct 2010
Make it a date night w The Romantics feat **Katie Holmes** & Josh Duhamel, now playing in Boston, SF, LA & NY: http://clicky.me/romanticstrailer
Expand Reply Retweet Favorite More

Dataset Creation

	Education	Job	Spouse
#Users	7,208	1,806	1,636
#Users Connected	6,295	1,407	1,108
#Edges	11,167	3,565	554
#Pos Entities	451	380	3121
#PosTweets	124,801	65,031	135,466
#Aver Pos Tweets User	17.3	36.6	82.8
#Neg Entity	6,987,186	4,405,530	8,840,722
#Neg Tweets	16,150,600	10,687,403	12,872,695

Table 1 : Statistics for our Dataset

Algorithm

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

Potential Function

Given an entity e recognized by Twitter NER (Ritter et al., 2011).

$\Psi(z_{i,e})$: Potential function, entity e constitutes the correspondent attribute of user i

$$\Psi(z_{i,e}) = \frac{1}{Z}\Psi_{Text}(z_{i,e})\Psi_{Network}(z_{i,e})$$

Learning

- Text-Level Evidence $\Psi_{Text}(z_{i,e}^k)$
  - Entity-level: number of tokens, capital letter, length
  - Token-level: identity, shape, POS, NER
  - Window-level: tokens, POS
  - Tweet-level: tokens
  - External Sources: list of universities/companies
- Neighboring Effect

Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)

$$\Psi_{Network}(z_{i,e}) = \prod_{j \in Neigh(i)} \exp(\lambda \mathbf{I}(Z_{j,e} = 1)/N)$$

  - Spouse

$$\Psi_{Network}(z_{i,e}) = \exp(\lambda \mathbf{I}(Z_{e,user_i} = 1))$$

- Max-Ent for training.

Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)

$$\Psi_{Network}(z_{i,e}) = \prod_{j \in Neigh(i)} \exp(\lambda \mathbf{I}(Z_{j,e} = 1)/N)$$

  - Spouse

$$\Psi_{Network}(z_{i,e}) = \exp(\lambda \mathbf{I}(Z_{e,user_i} = 1))$$

- Max-Ent for training.

Learning

- Text-Level Evidence $\Psi_{Text}(z^k_{i,e})$
- Neighboring Effect
  - Education, Job (Homophily)

$$\Psi_{Network}(z_{i,e}) = \prod_{j \in Neigh(i)} \exp(\lambda \mathbf{I}(Z_{j,e} = 1)/N)$$

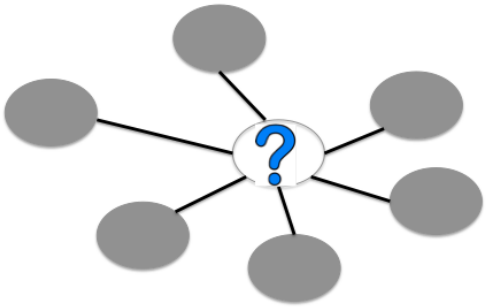
  - Spouse

$$\Psi_{Network}(z_{i,e}) = \exp(\lambda \mathbf{I}(Z_{e,user_i} = 1))$$

- Max-Ent for training.

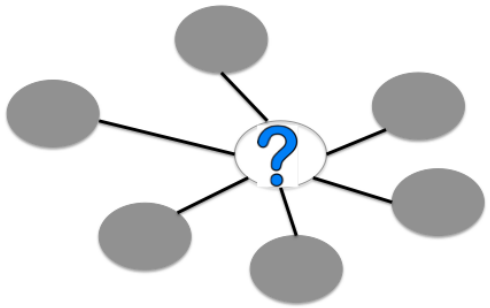
Inference

- **Observed**: Neighboring Information is already given (Education, Job).

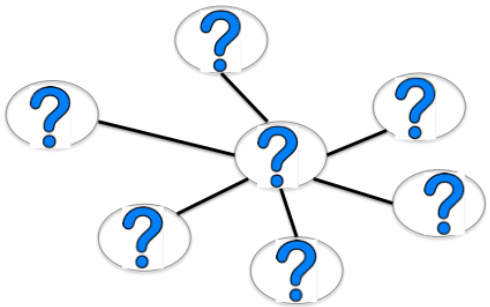


Inference

- **Observed**: Neighboring Information is already given (Education, Job).

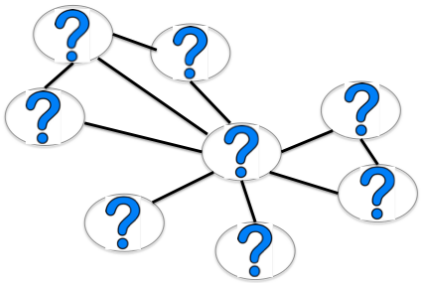


- **Latent**: Neighboring Information is unknown (Joint Inference)



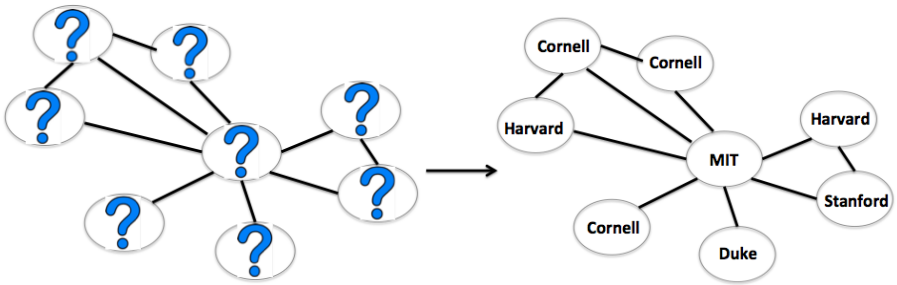
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**



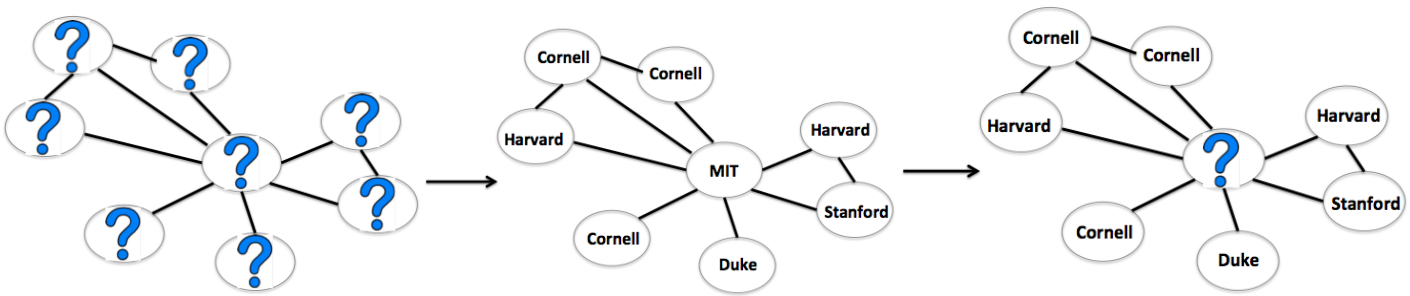
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - **Initializing only based on text-level information $\Psi_{Text}(z_{i,e})$**



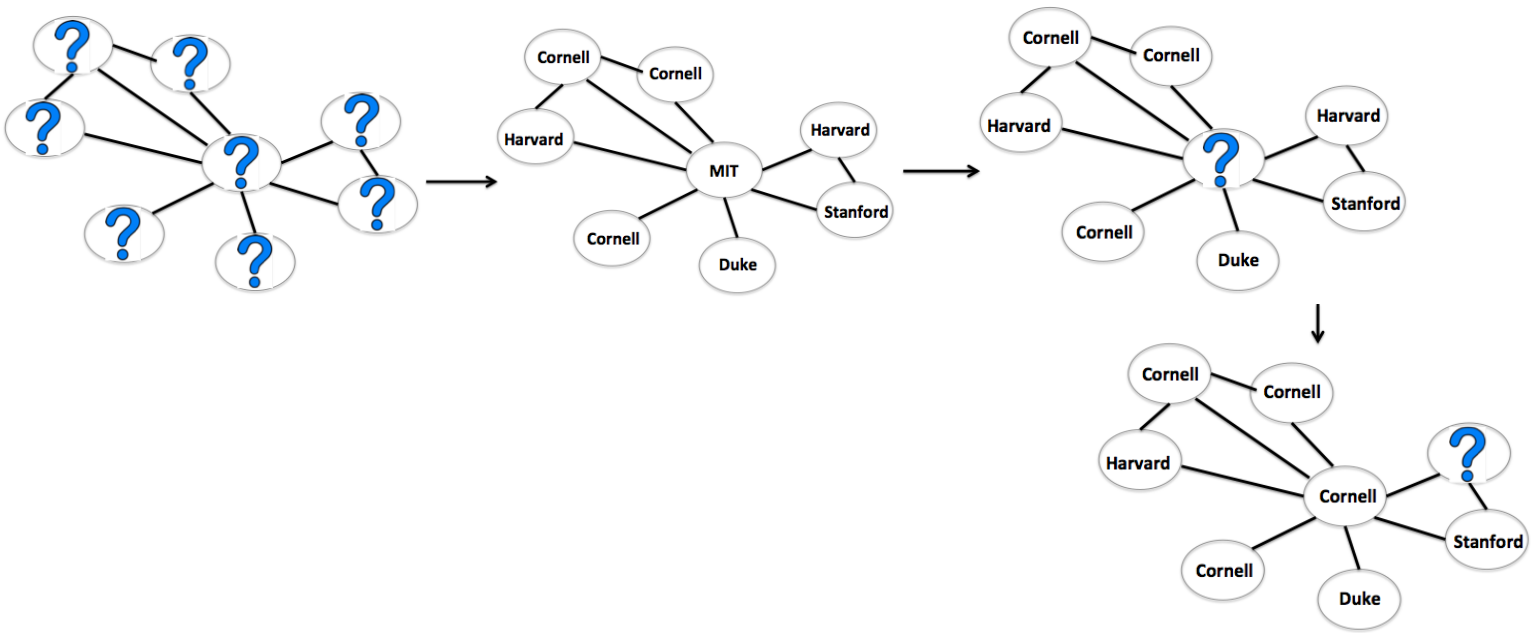
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - **Initializing only based on text-level information $\Psi_{Text}(z_{i,e})$**
  - **Infer each individual given its neighbors**



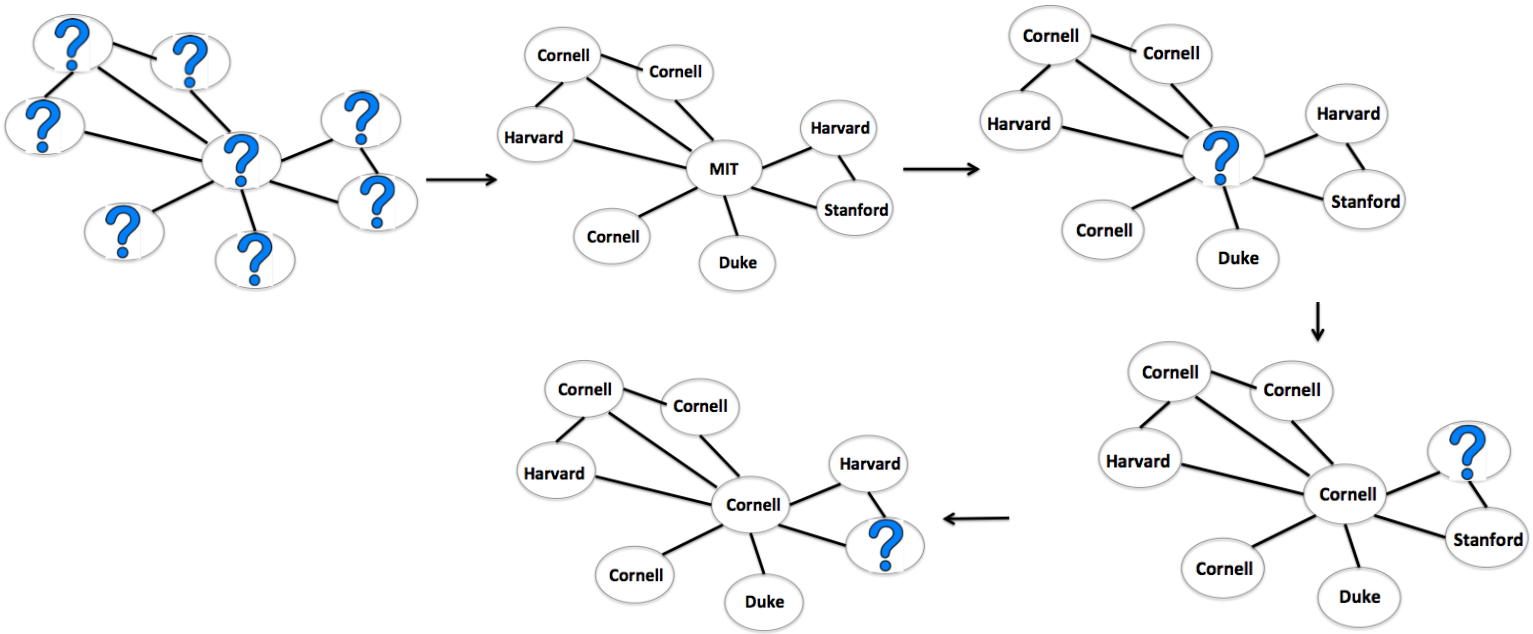
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - **Initializing only based on text-level information $\Psi_{Text}(z_{i,e})$**
  - **Infer each individual given its neighbors**



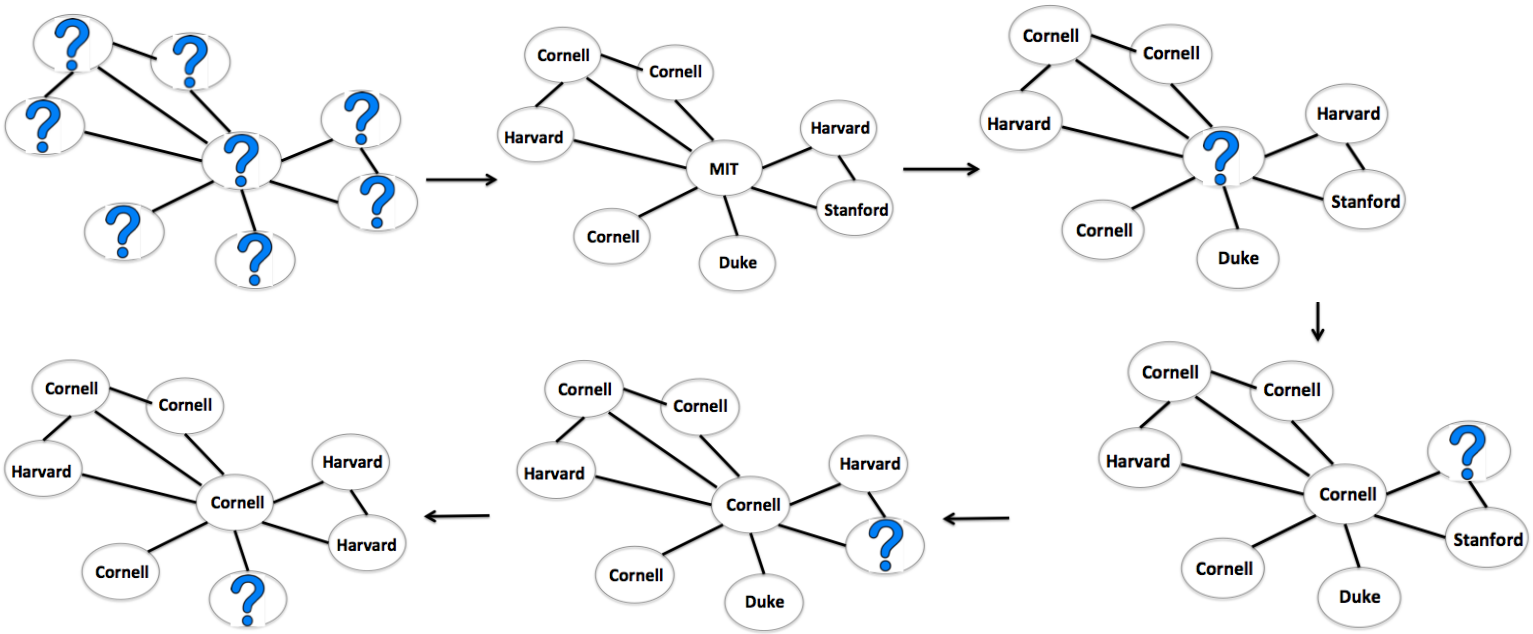
Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - **Initializing only based on text-level information $\Psi_{Text}(z_{i,e})$**
  - **Infer each individual given its neighbors**



Inference

- **Latent: Neighboring Information is unknown (Joint Inference)**
  - **Initializing only based on text-level information $\Psi_{Text}(z_{i,e})$**
  - **Infer each individual given its neighbors**



Experiments

- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

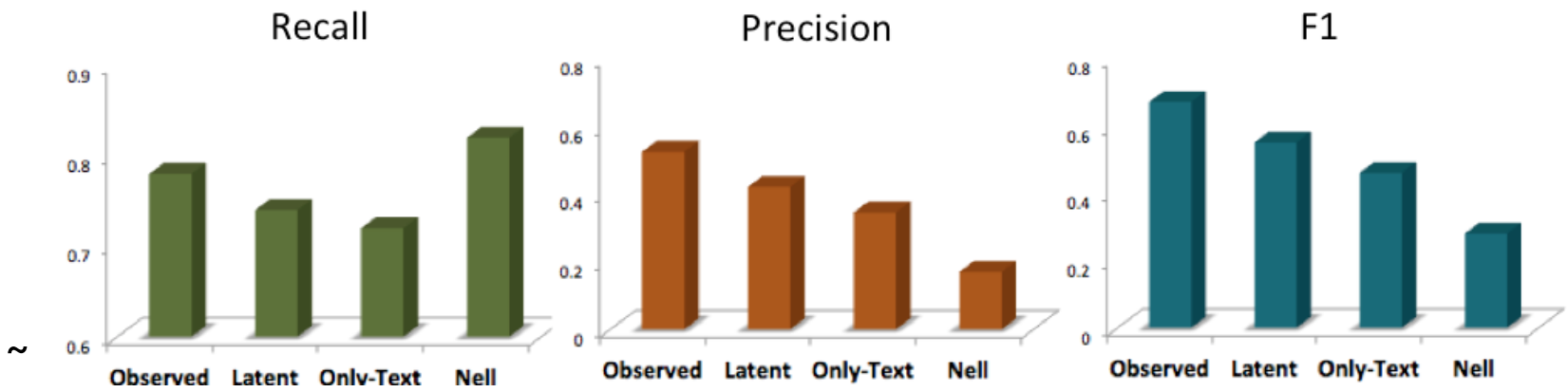
Baselines

- Only-Text:
  Text-Level Evidence $\Psi_{Text}(z_{i,e})$

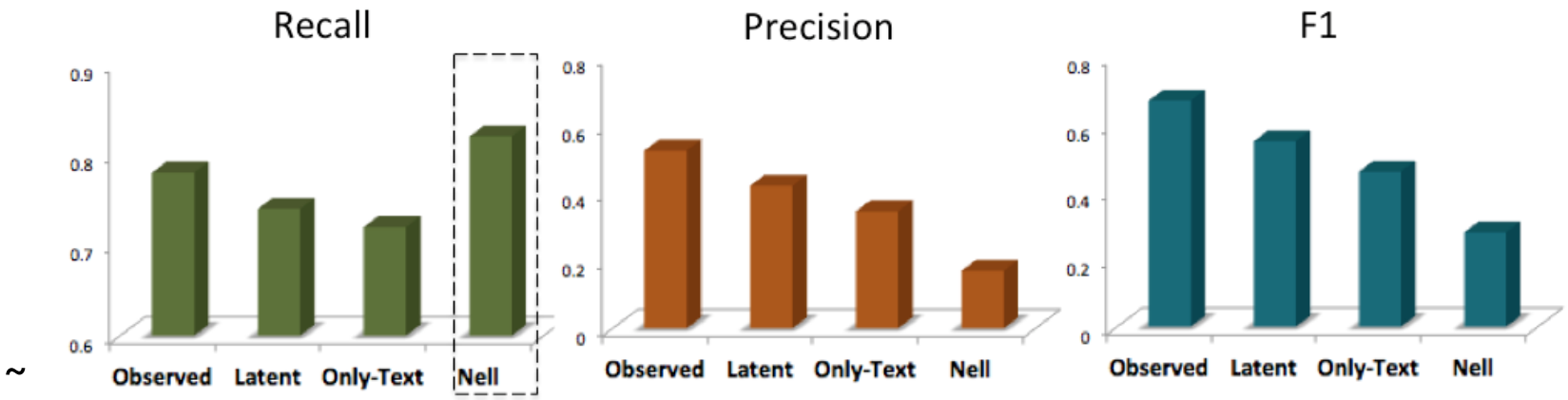
Baselines

- Only-Text:
  Text-Level Evidence $\Psi_{Text}(z_{i,e})$
- NELL: Bag of words matching in the list of universities or companies borrowed from NELL

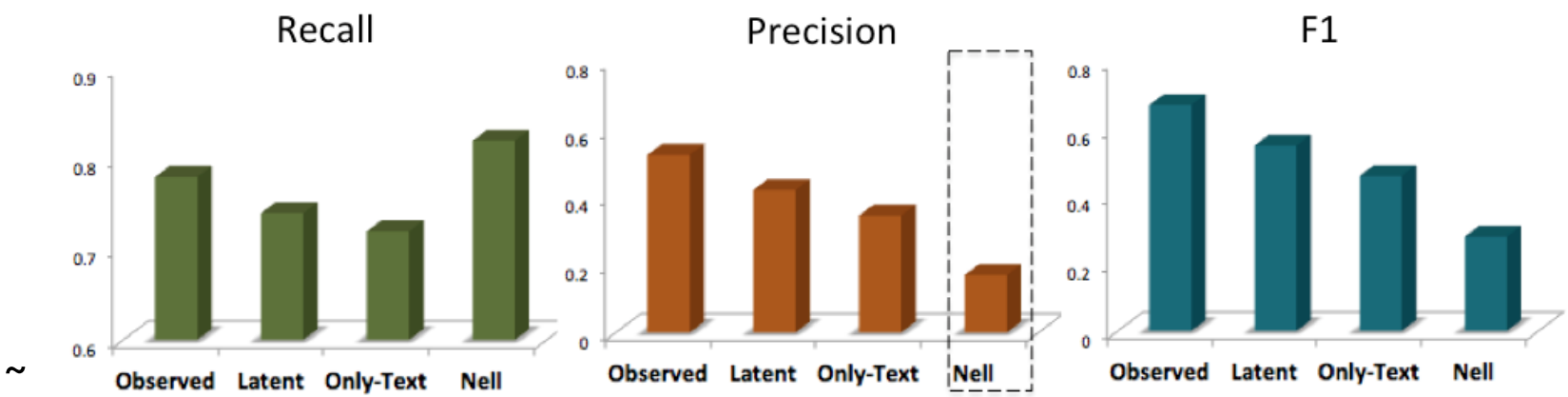
Results


Recall
0.9
0.8
0.7
0.6
Observed
Latent
Only-Text
Nell
Precision
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell
F1
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell

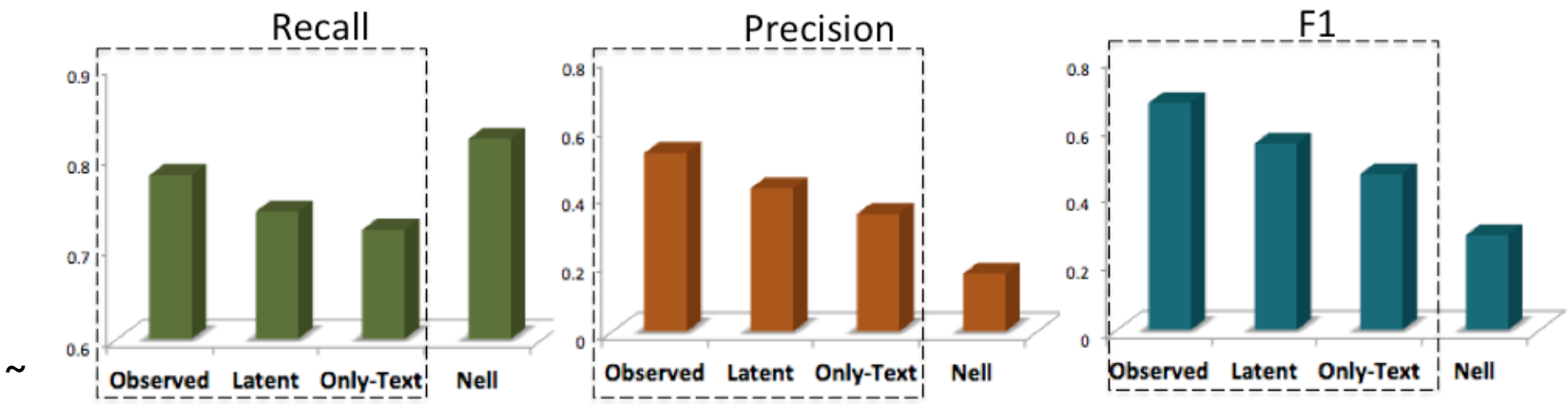
Results


Recall
0.9
0.8
0.7
0.6
Observed
Latent
Only-Text
Nell
Precision
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell
F1
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell

Results


Recall
0.9
0.8
0.7
0.6
Observed
Latent
Only-Text
Nell
Precision
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell
F1
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell

Results


Recall
0.9
0.8
0.7
0.6
Observed
Latent
Only-Text
Nell
Precision
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell
F1
0.8
0.6
0.4
0.2
0
Observed
Latent
Only-Text
Nell

Conclusion

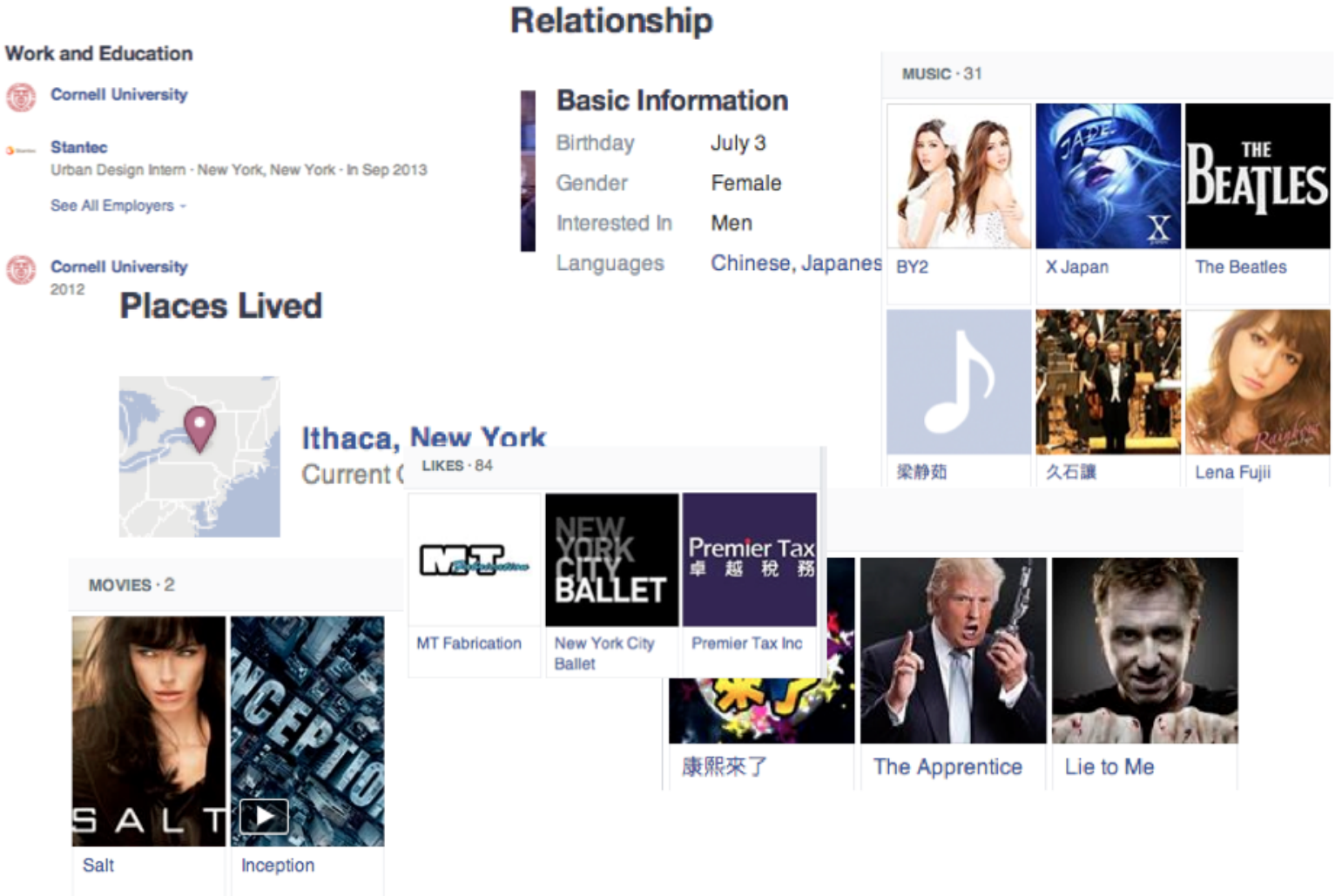
- Motivation/Introduction
- Related Work
- Dataset Creation
- Algorithm
- Experiments
- Conclusion

Conclusion

- We present a framework to extract user attributes from Twitter.
- We present a large-scale dataset for this task.
- We demonstrate the benefit of jointly reasoning about users' social network structure.

Future Work

Facebook:



Thank you

Dataset

http://aclweb.org/aclwiki/title=Profile_data

Thank you

Dataset

`http://aclweb.org/aclwiki/title=Profile_data`

Thank You !

Questions, Suggestions ?