

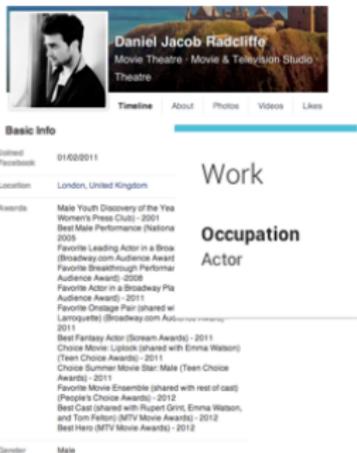
Weakly Supervised User Profile Extraction from Twitter

Jiwei Li, Alan Ritter and Eduard Hovy

School of Computer Science, Carnegie Mellon University

June 22nd, 2014

User Profile



Daniel Jacob Radcliffe
Movie Theatre · Movie & Television Studio
Theatre

Timeline About Photos Videos Likes

Basic Info

Joined Facebook: 01/03/2011

Location: London, United Kingdom

Work

Occupation Actor

Awards

- Male Youth Discovery of the Year (Women's Press Club) - 2001
- Best Male Performance (National) - 2005
- Favorite Leading Actor in a Block (Broadway.com Audience Award)
- Favorite Breakthrough Performer (Audience Award) - 2008
- Favorite Actor in a Broadway Play (Audience Award) - 2011
- Favorite Onstage Pair (shared w/ Lemony Snicket) (Broadway.com Audience Award) - 2011
- Best Fantasy Actor (Stream Awards) - 2011
- Choice Movie: Lemony (shared with Emma Watson) (Teen Choice Awards) - 2011
- Choice Summer Movie Star: Male (Teen Choice Awards) - 2011
- Favorite Movie Ensemble (shared with rest of cast) (People's Choice Awards) - 2012
- Best Cast (shared with: Rupert Grint, Emma Watson, and Tom Felton) (SMTV Movie Awards) - 2012
- Best Hero (MTV Movie Awards) - 2012

Gender: Male



Education



Gender



Spouse/Lover



Religion



Job

Interest
Favorite Movie

Interest



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 Beograd

Class of 1999 · Ljubljana, Slovenia

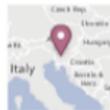


Places Lived



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




Fernando Pereira

Works at Google
Attended University of Lisbon
Lives in Palo Alto, California

Following

3,834 have him in circles · 10 in common

🌐 | 🗨️ | ⌵

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 ...)
-



Already in a relationship with someone ?

Emma Watson [@EmWatson](#)
British actress
London
[emmawatson.com](#)
Joined July 2010

TWEETS 716 PHOTOS/VIDEOS 67 FOLLOWING 95 FOLLOWERS 13.3M FAVORITES 342 More v

Tweets Tweets and replies

Retweeted by Emma Watson

Derek Blasberg [@DerekBlasberg](#) · Jun 1
Wait, what? it's June already?

Motivation/Introduction



Profile summary



Bill Simmons ✓
@BillSimmons

Grantland boss • columnist, @30for30 co-creator, NBA Countdown co-host, BS Report host, author of thebookofbasketball.com, On Facebook: facebook.com/billsimmons
Los Angeles (via Boston) - grantland.com

Followed by NHL and NBA.

Bill Simmons @BillSimmons · 10m
To repeat: Flip Saunders has more power/equity/say than anyone else in the NBA - even Popovich, Doc, Spo + SVG! Flip Saunders!!!! WTF?????

Details

Bill Simmons @BillSimmons · 13m
Flip Saunders is a head coach/team prez/part-owner??? Kudos to Glen Taylor - with Sterling gone, he's officially the NBA's worst owner. Wow.

Details

Profile summary



Kenny Smith ✓
@TheJetOnTNT

Championships and Emmy's, Kenny Smith
The Universe - kennythejetsmith.com

Followed by NBA.

Kenny Smith @TheJetOnTNT · Jun 3
Guess who mrsgwenniesmith and I ran into today... #adayintheife
[instagram.com/prozogie03/](https://www.instagram.com/prozogie03/)

Details

Kenny Smith @TheJetOnTNT · Jun 3
For a team to succeed, you need chemistry. Here is a look at the teams that I believe are doing it the best bit.ly/19AZa5F

Details

Motivation/Introduction

Twitter serves as a wonderful source:

Motivation/Introduction

Twitter serves as a wonderful source:

- Text Level Evidence

 **Omar Rasheed**  
@OmarRasheed503
HARVARD ACCEPTED ME 🙏🙏🙏
pic.twitter.com/VPsRo97VzE

← Reply ↻ Retweet ★ Favorite ... More

 **Ana García Puyol**  
@anagpuyol
It's been an honor studying at **@Harvard** for two years and going to class with really talented people who I can now call friends.

← Reply ↻ Retweet ★ Favorite ... More

11:01 PM - 29 May 2014

 **Ana García Puyol**  
@anagpuyol
Think! #harvard14 #graduation **@Harvard** University [instagram.com/p/olfB_1jCEg/](https://www.instagram.com/p/olfB_1jCEg/)

📍 From Cambridge, MA

← Reply ↻ Retweet ★ Favorite ... More

8:25 AM - 29 May 2014

Motivation/Introduction

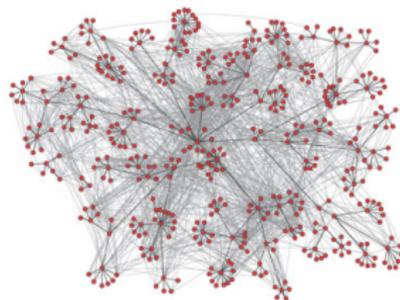
Twitter serves as a wonderful source:

- 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 → 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)

 Freebase

IsCapital (Paris, France)

IsCapital (London, Britain)



Dataset Creation

Distant Supervision

- Relation Extraction (Mintz et al., 2009)

 Freebase

IsCapital (Paris, France)

IsCapital (London, Britain)



- 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)

 Freebase

IsCapital (Paris, France)

IsCapital (London, Britain)



- 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 ?

The image shows a user profile for Fernando Pereira. The profile includes a circular profile picture, a name, and a bio: "Works at Google, Attended University of Lisbon, Lives in Palo Alto, California". Below the bio is a "Work" section listing his employment history: Google (Research director, 2008 - present), LNEC, Lisbon (1975 - 1977), SRI International (1982 - 1989), AT&T (1989 - 2000), and University of Pennsylvania (2001 - 2007). There is also an "Education" section listing the University of Lisbon (Mathematics, 1975 - 1977) and the University of Pennsylvania (Artificial Intelligence, 2001 - 2007). A "Links" section includes a Google+ URL, a YouTube channel, and other profiles like Blogger, Google Reader, and Picasa Web Albums. A tweet from @eammyturns is shown below, with a red box around the text "Google's Hybrid Approach to Research | July 2012 | Communications of the ACM" and an arrow pointing from the "Google" entry in the work history to this tweet. Another red box is around the Twitter handle "@eammyturns" in the "Other profiles" section, with an arrow pointing to the tweet's author name. The tweet has 6 retweets and 3 favorites.

Dataset Creation

Attributes we focus on:

- Education
- Job
- Spouse

Dataset Creation

Attributes we focus on:

- Education
- Job
- Spouse

Dataset Creation

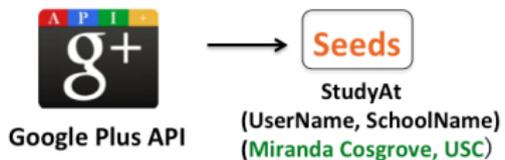
- Education: Positive Examples



Google Plus API

Dataset Creation

- Education: Positive Examples



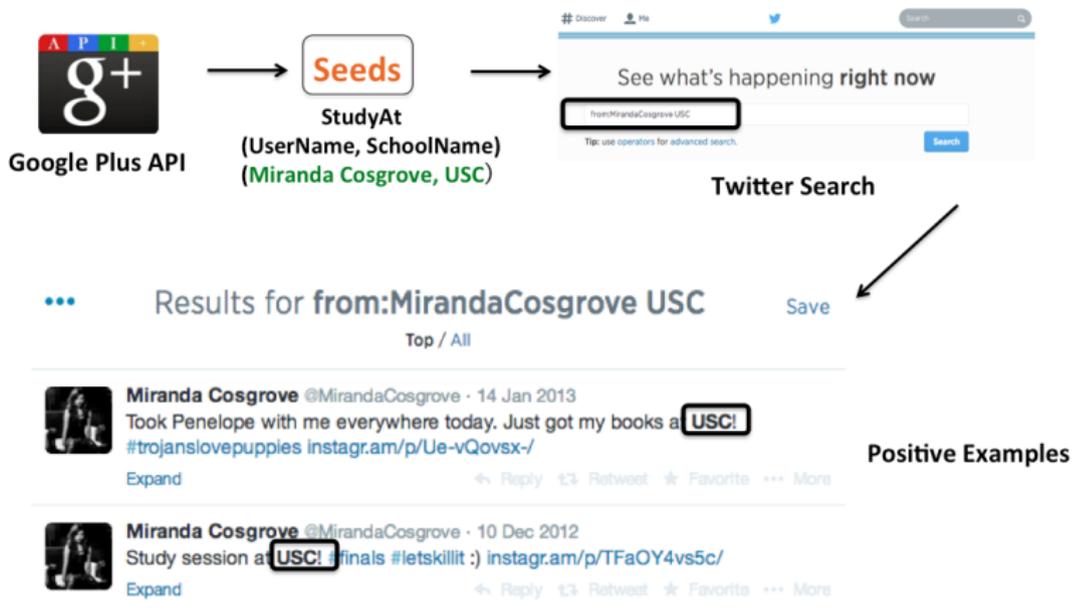
Dataset Creation

- Education: Positive Examples



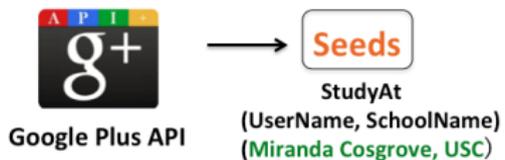
Dataset Creation

- Education: Positive Examples



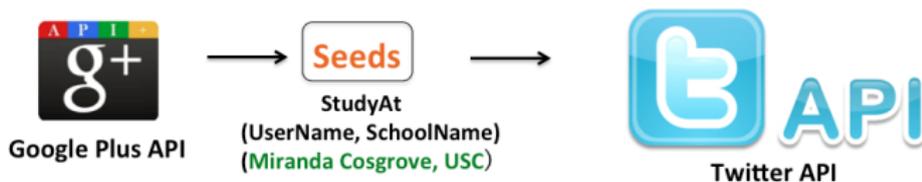
Dataset Creation

- Education: Negative Examples



Dataset Creation

- Education: Negative Examples



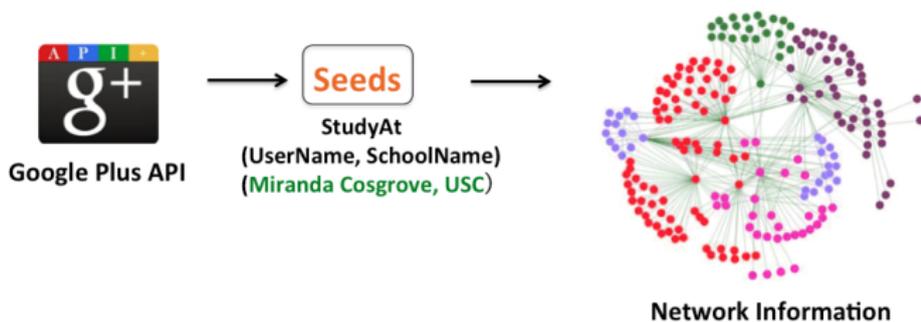
Dataset Creation

- Education: Negative Examples



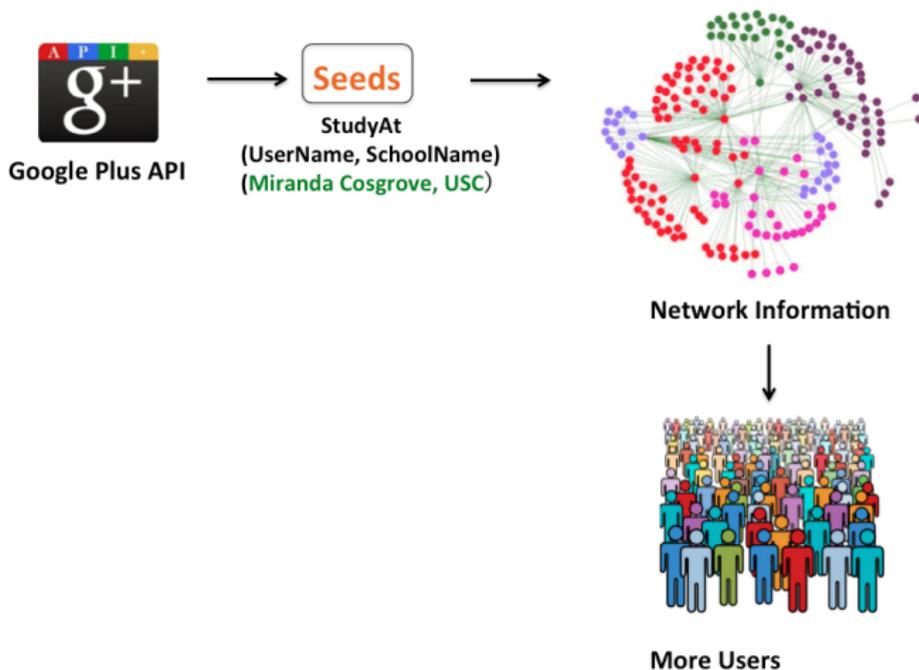
Dataset Creation

- Education: Data Expansion



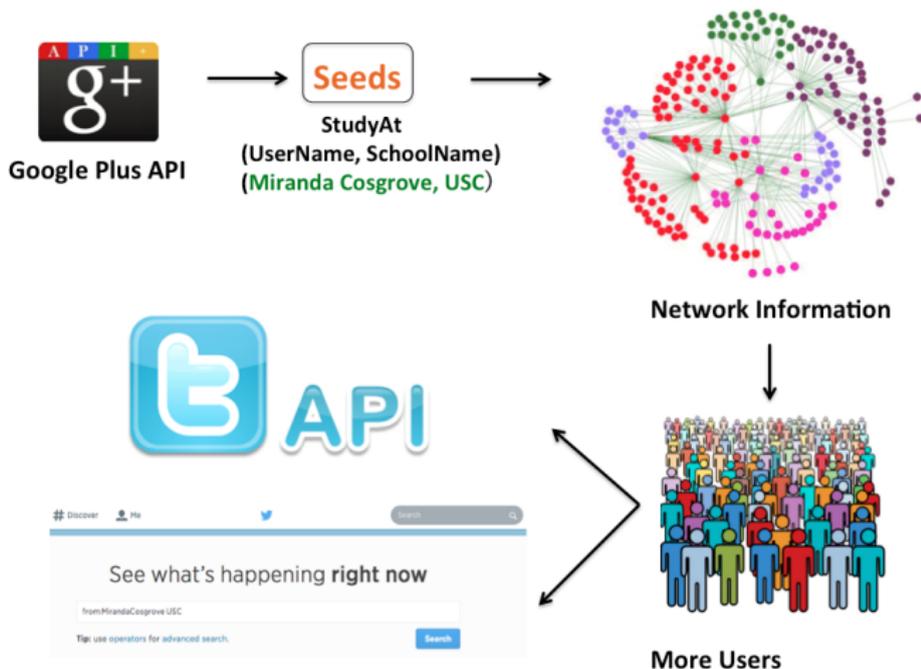
Dataset Creation

- Education: Data Expansion



Dataset Creation

- Education: Data Expansion



Dataset Creation

- Spouse



Freebase API

Dataset Creation

- Spouse



Freebase API

IsSpouse
(Tom Cruise, Katie Holmes)



Dataset Creation

- Spouse



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
#Pos Tweets	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 $\Psi_{Text}(z_{i,e}^k)$
- 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 $\Psi_{Text}(z_{i,e}^k)$
- 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 $\Psi_{Text}(z_{i,e}^k)$
- Neighboring Effect
 - Education, Job (Homophily)

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

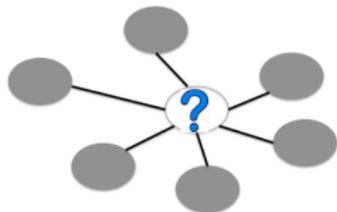
- Spouse

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

- Max-Ent for training.

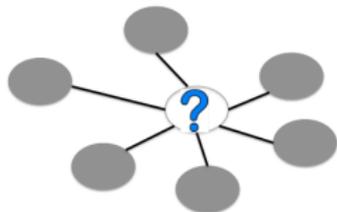
Inference

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



Inference

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



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



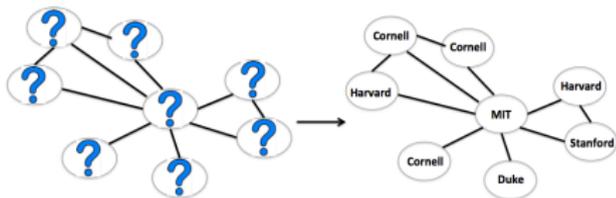
Inference

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



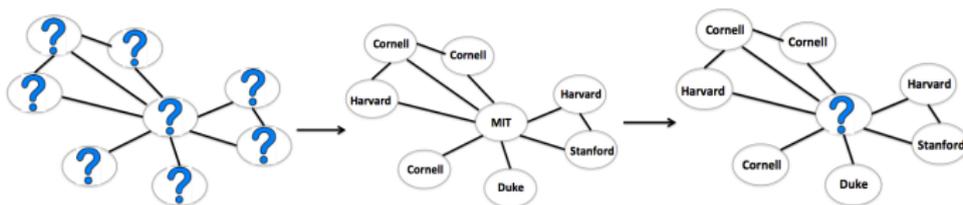
Inference

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



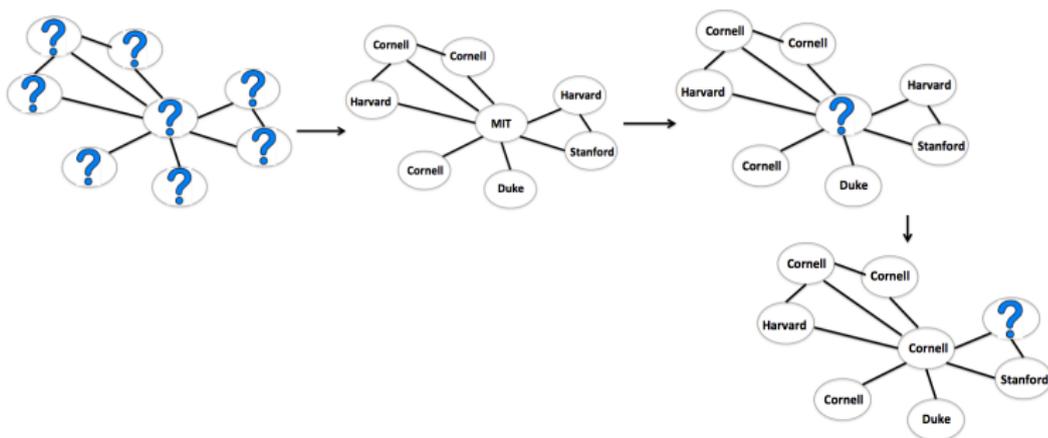
Inference

- **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



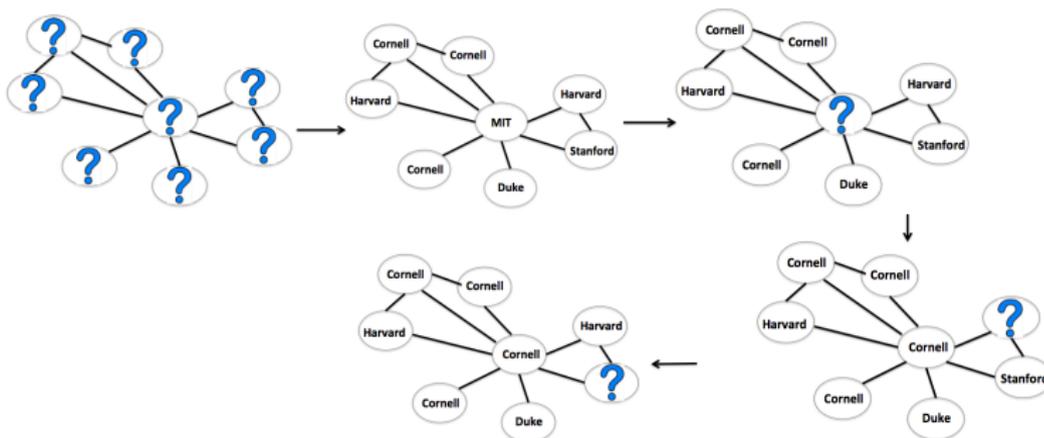
Inference

- **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



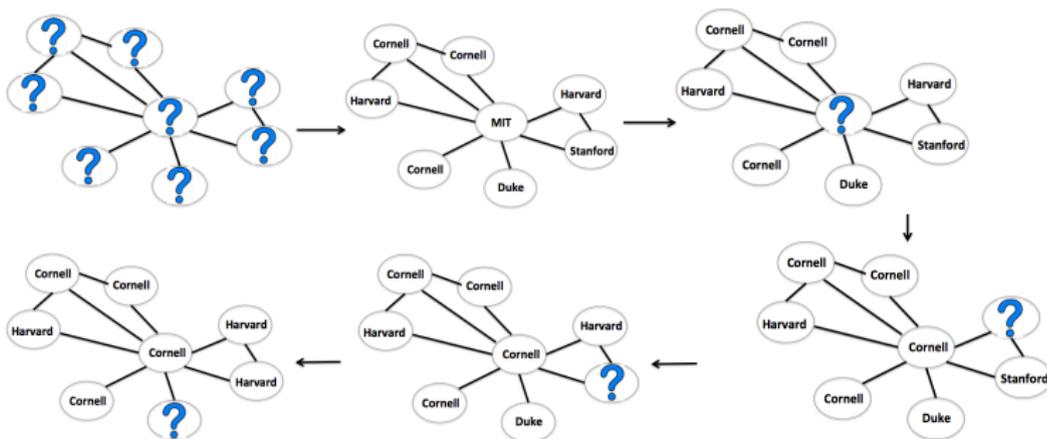
Inference

- **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

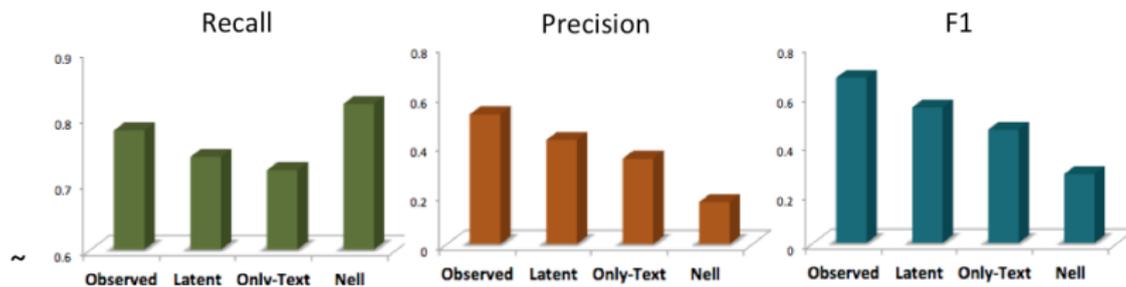
Baselines

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

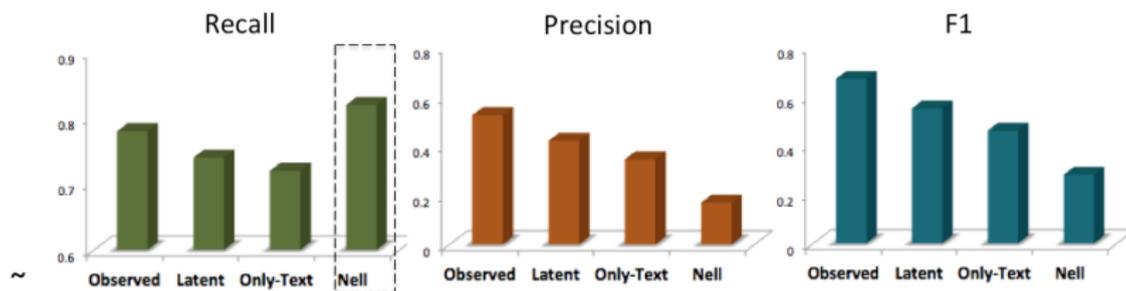
Baselines

- 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

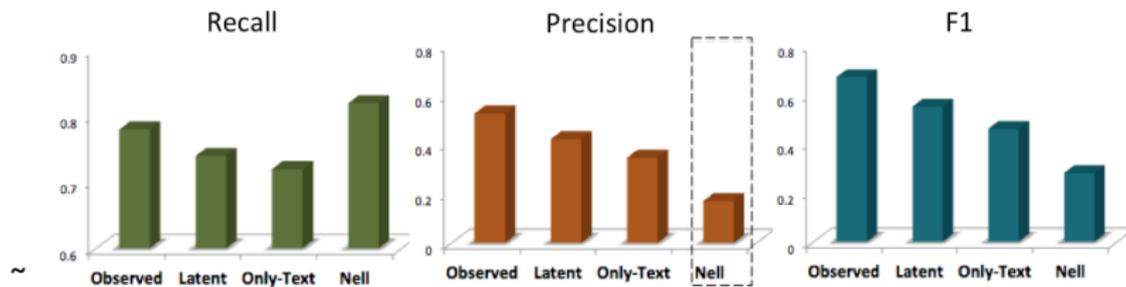
Results



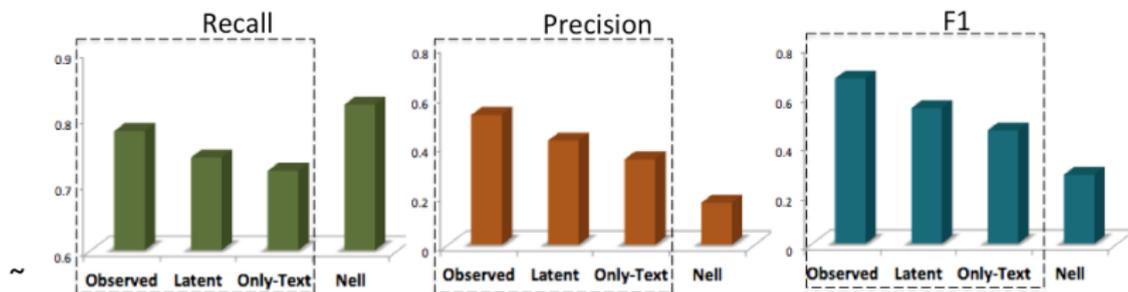
Results



Results



Results



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:

Work and Education

-  **Cornell University**
-  **Stanec**
Urban Design Intern - New York, New York - In Sep 2013
[See All Employers -](#)
-  **Cornell University**
2012

Places Lived



Ithaca, New York

Current (

LIKES - 84

MOVIES · 2



Salt



Inception

Relationship

Basic Information

Birthday July 3
 Gender Female
 Interested In Men
 Languages Chinese, Japanese

MUSIC · 31



BY2



X Japan



The Beatles



梁静茹



久石讓



Lena Fujii



MT Fabrication



New York City Ballet



Premier Tax Inc



康熙來了



The Apprentice



Lie to Me

Thank you

Dataset

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

Thank you

Dataset

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

Thank You !

Questions, Suggestions ?