Web Science – **Investigating the Future** of Information and Communication

#science

Identifying and Analyzing Researchers on Twitter

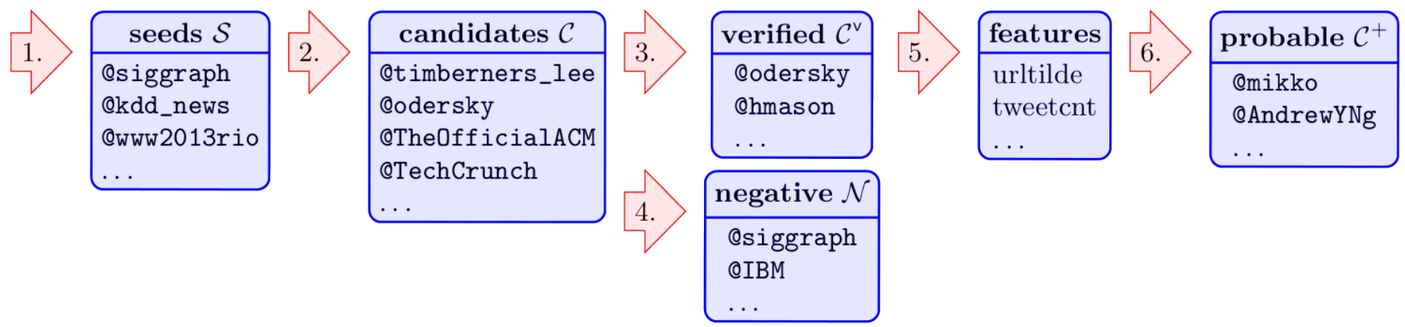
Robert Jäschke

Asmelash Teka Hadgu

Agenda

- **Motivation**

- Approach



- Results

Twitter is …

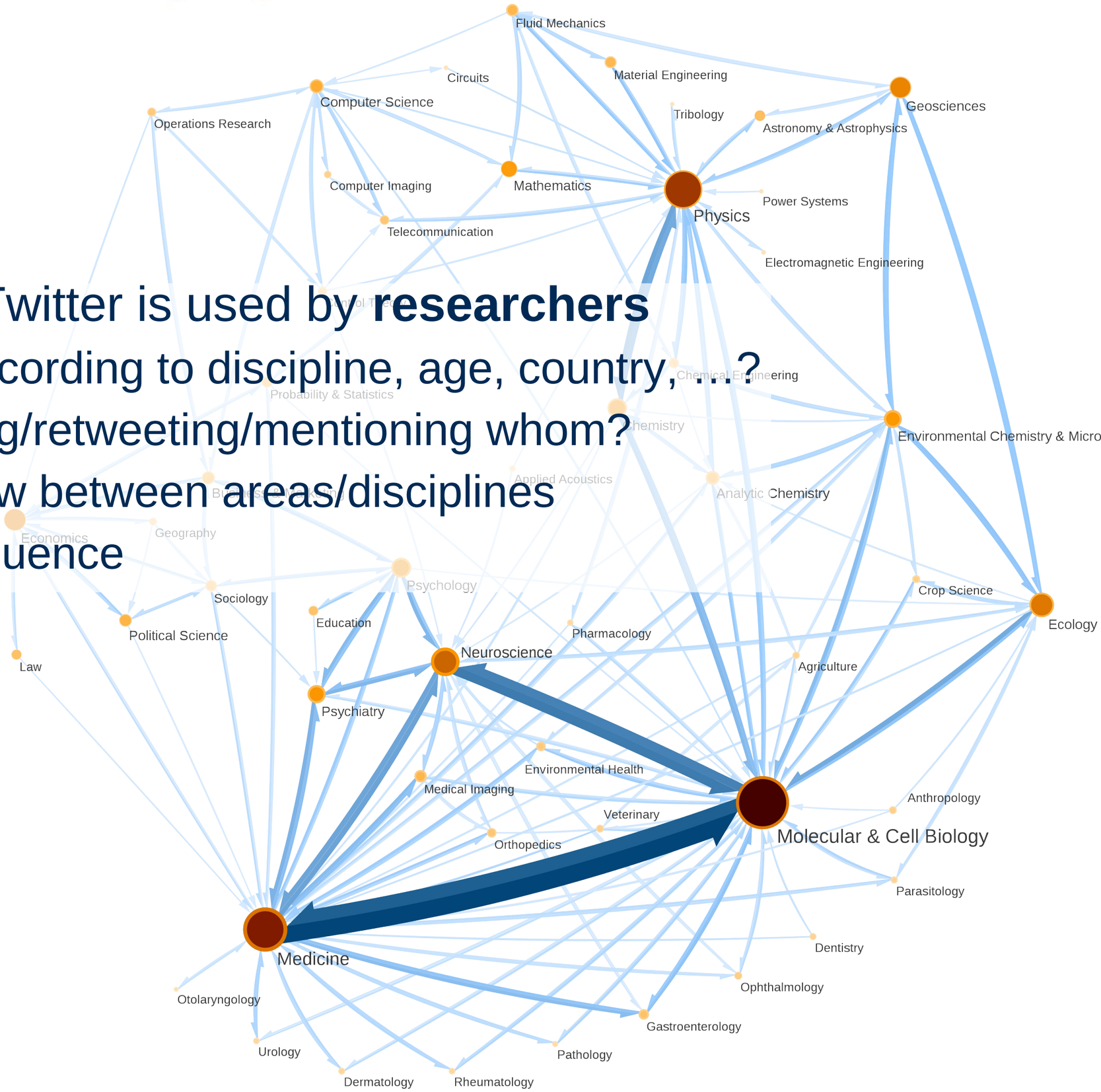
- a communication platform,
- a social network,
- a system for resource sharing

… which **researchers** use …

- to connect with other researchers,
- to announce calls for papers,
- to communicate and discuss,
- to stay up-to-date,
- etc.

Goals

- Understand how Twitter is used by **researchers**
  - Differences according to discipline, age, country, ...?
  - Who's following/retweeting/mentioning whom?
  - Information flow between areas/disciplines
  - Impact and influence



Goals

- Improve retrieval and discovery of **scientific content**
  - Researchers, topics, publications, conferences, …
  - Trends, developments over time
  - Personalized recommendations

Goals

- Transfer **peer review** to social media
  - What do *researchers* regard as important?

Challenges

- Data acquisition
  - Tweets and users from Twitter
  - Ground truth to train and evaluate algorithms
- Identifying researchers
  - One class problem: finding good counterexamples is difficult
- Brevity of tweets
  - How to extract meaning from 140 characters?
- Identifying and classifying scientific tweets
  - What is a scientific tweet?
- Ranking scientific content
  - How to evaluate a ranking?

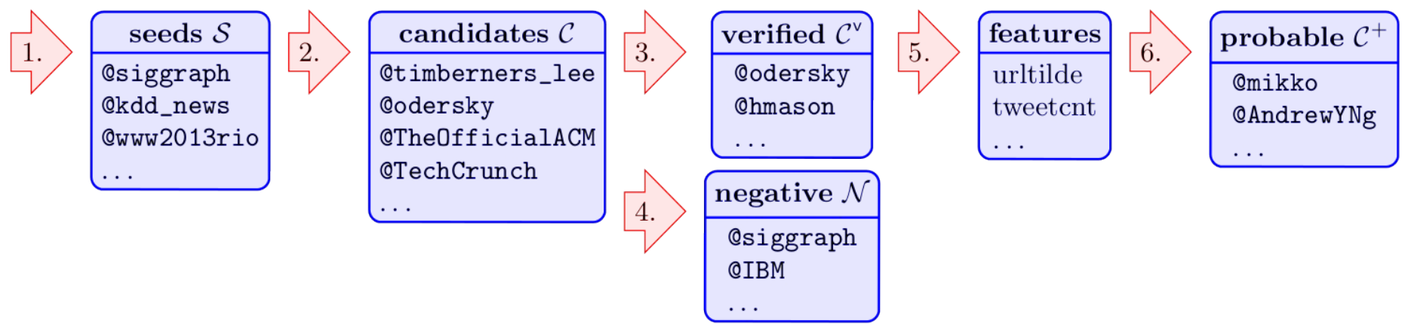
Related Work

- Twitter directories (e.g., Wefollow, Twellow, JustTweetIt)
- User classification:
  - D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in Twitter (2010)
  - M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in Twitter (2011)
- Scholars on Twitter:
  - J. Priem and B. Hemminger. Scientometrics 2.0: New metrics of scholarly impact on the social web (2010)
  - J. Letierce, A. Passant, J. Breslin, and S. Decker. Understanding how Twitter is used to widely spread scientific messages (2010)
  - K. Weller, E. Dröge, and C. Puschmann. Citation analysis in Twitter: Approaches for defining and measuring information flows within tweets during scientific conferences (2011)
  - G. Eysenbach. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact (2011)
- Typically: focus on *tweets*, not *users*

Agenda

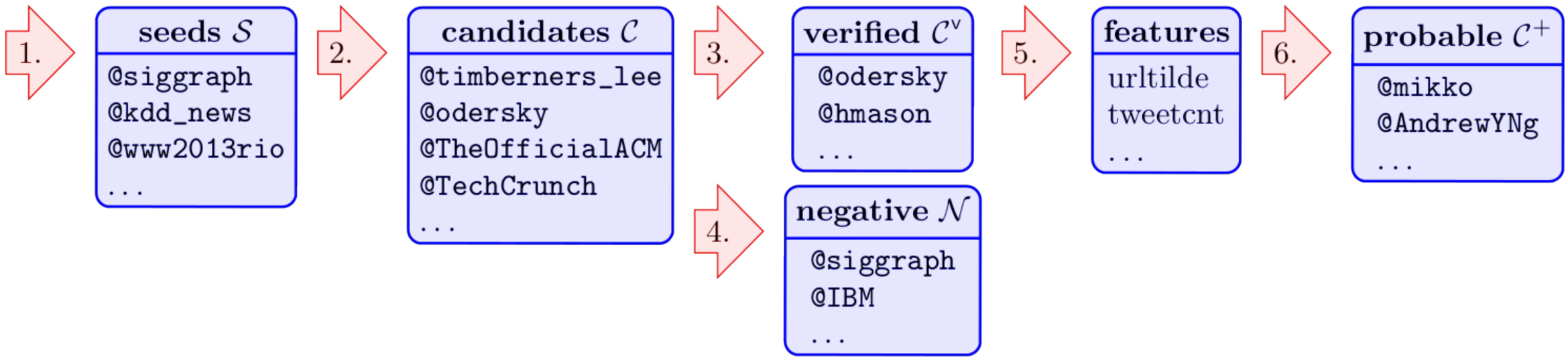
- Motivation

- **Approach**



- Results

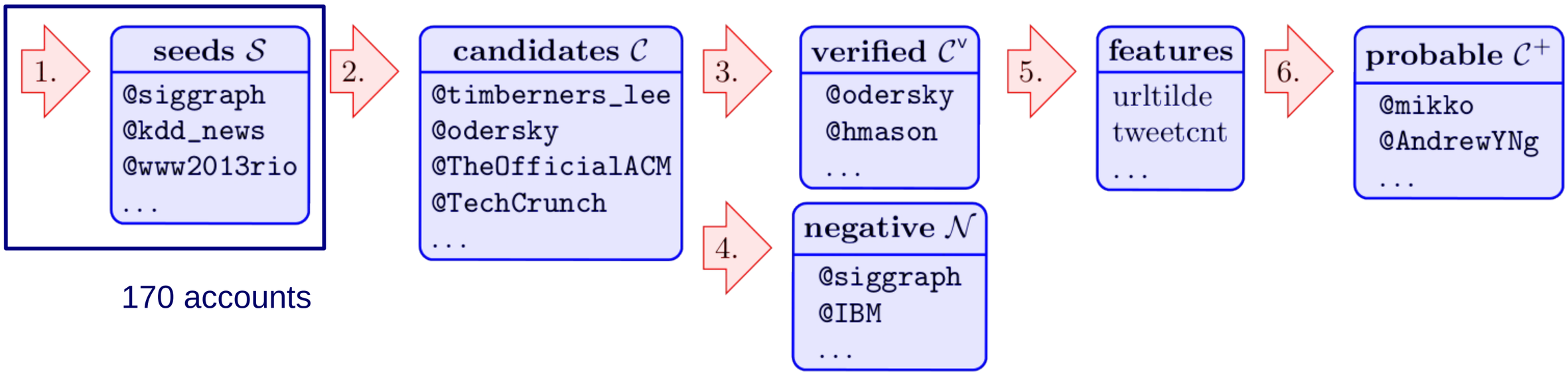
Approach



As a first step, we

- focused on c*omputer science*
- developed a pipeline to *identify* researchers
- *analyzed* their age, popularity, influence, and social network

Approach

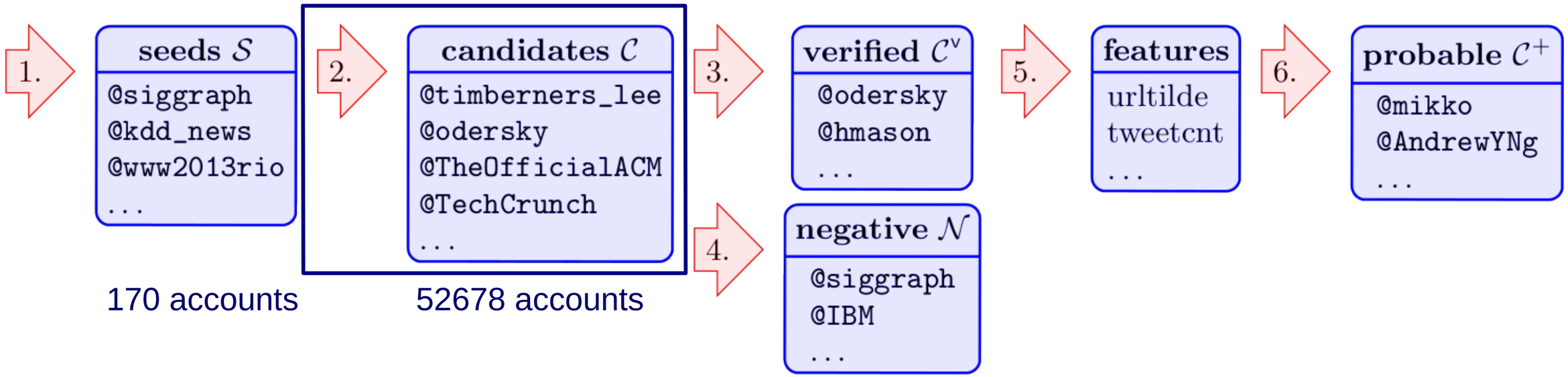


Finding good **seeds**:

- requirements: small set, good coverage, followers likely scientists
- solution: Twitter accounts of computer science conferences
- started with a list from Wikipedia[1], searched for Twitter accounts
- 170 accounts for 98 conferences

1: http://en.wikipedia.org/wiki/List_of_computer_science_conferences

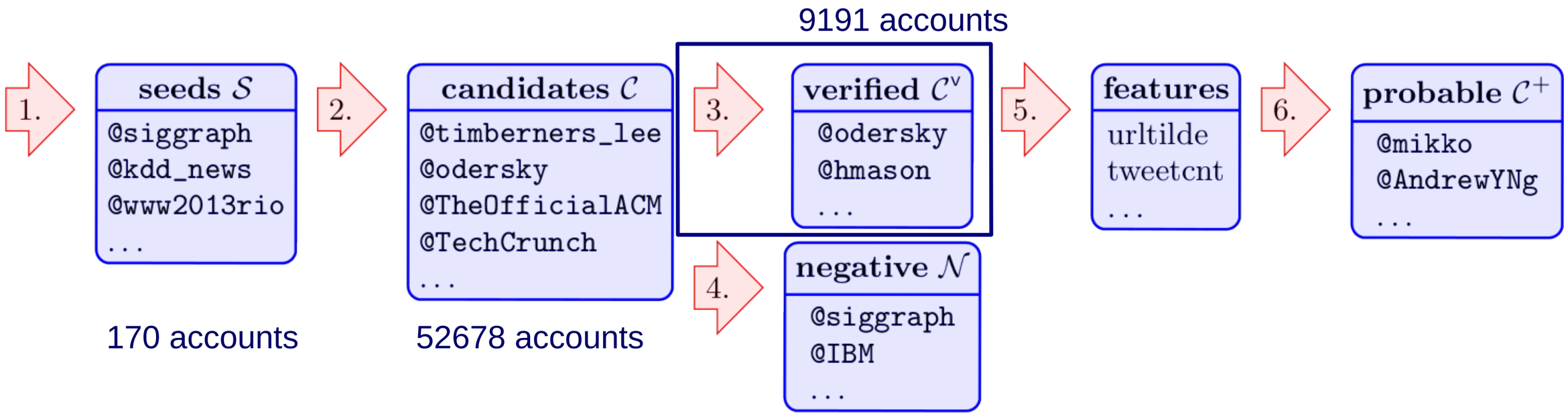
Approach



Generating **candidates**:

- follower, friends, retweeter of the seeds
- recursive approach possible but reduces precision
- 52678 accounts, mostly interested in one conference (83%)

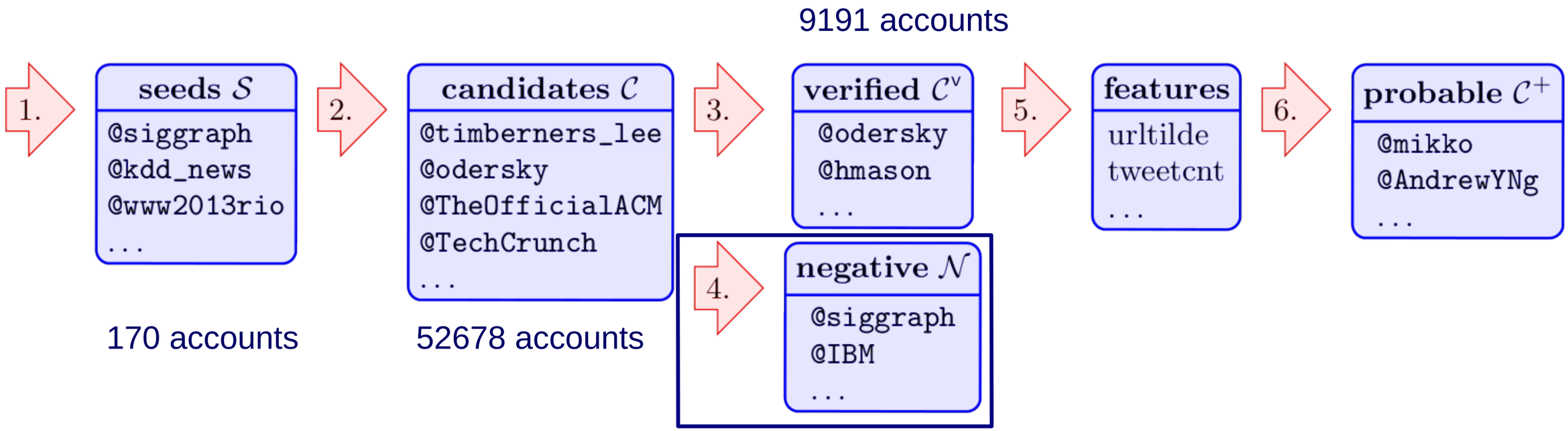
Approach



Verifying candidates with ground truth:

- using computer science publications as evidence
- matching against 1304283 author names from DBLP
- matching: string matching of real name, ignoring duplicates
- manual validation of 150 verified accounts: 73% accuracy

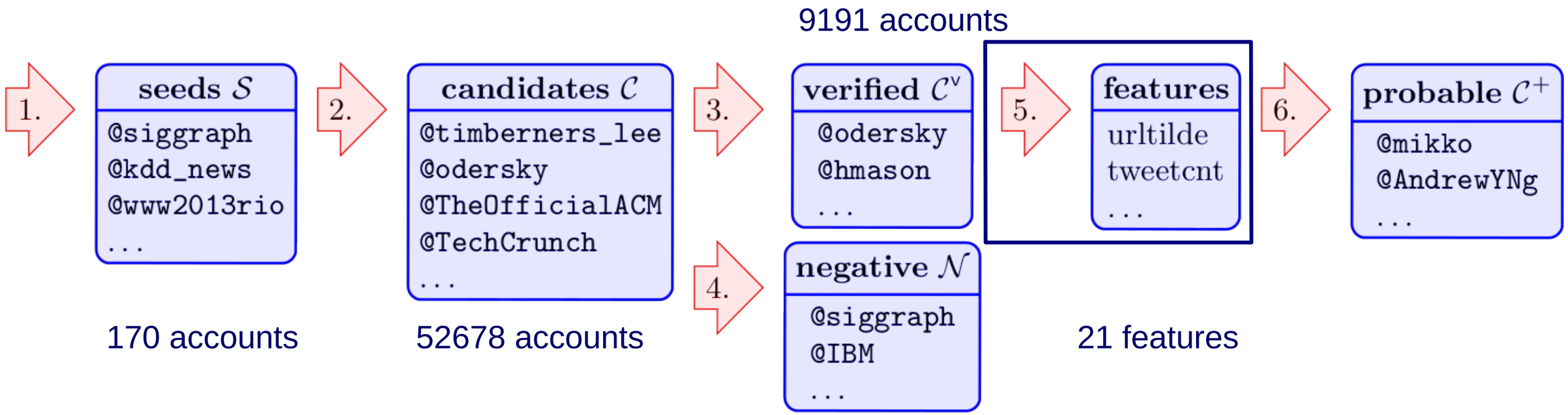
Approach



Finding **negative examples**:

- challenging task: most users are *not* researchers
  - How to get a representative sample?
- randomly collected users from the Twitter stream
- removed candidates, their followers and friends
- added seeds and large companies

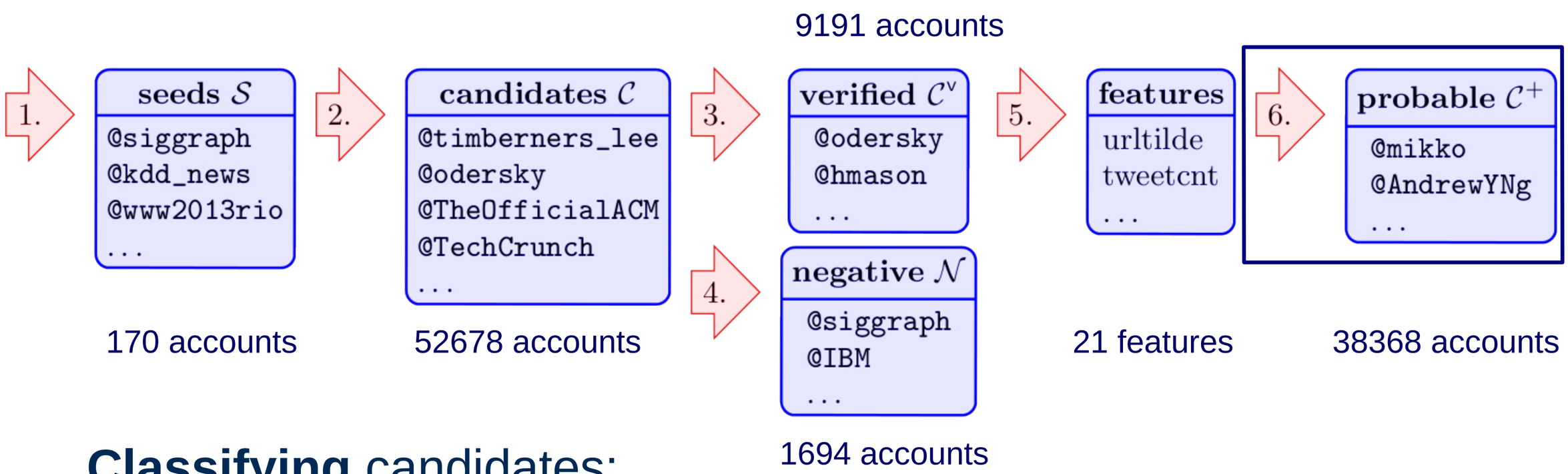
Approach



Extracting **features**:

- Which features can separate researchers from other users?
- *profile* (#tweets, #followers, website set, bio keywords, etc.) and
- *content* (#tweets with URLs, #scientific tweets, etc.) features,
- no *network* (#followed seeds, etc.) features

Approach



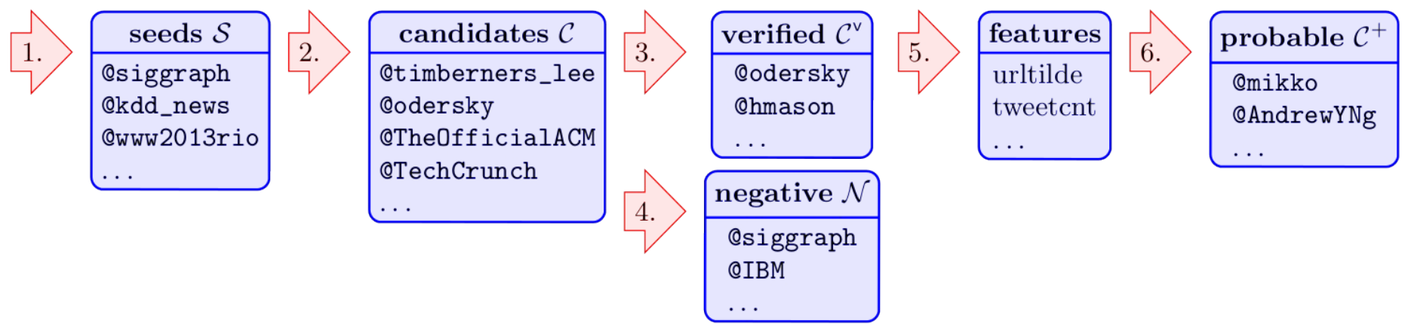
Classifying candidates:

- stratified 10-fold cross validation (2000 random cand. + neg. ex.)
- Random Forest: F1 of 0.94
- Baseline (SVM on Bag of Words): F1 of 0.93
- 38368 positive candidates, 5015 negative candidates

Agenda

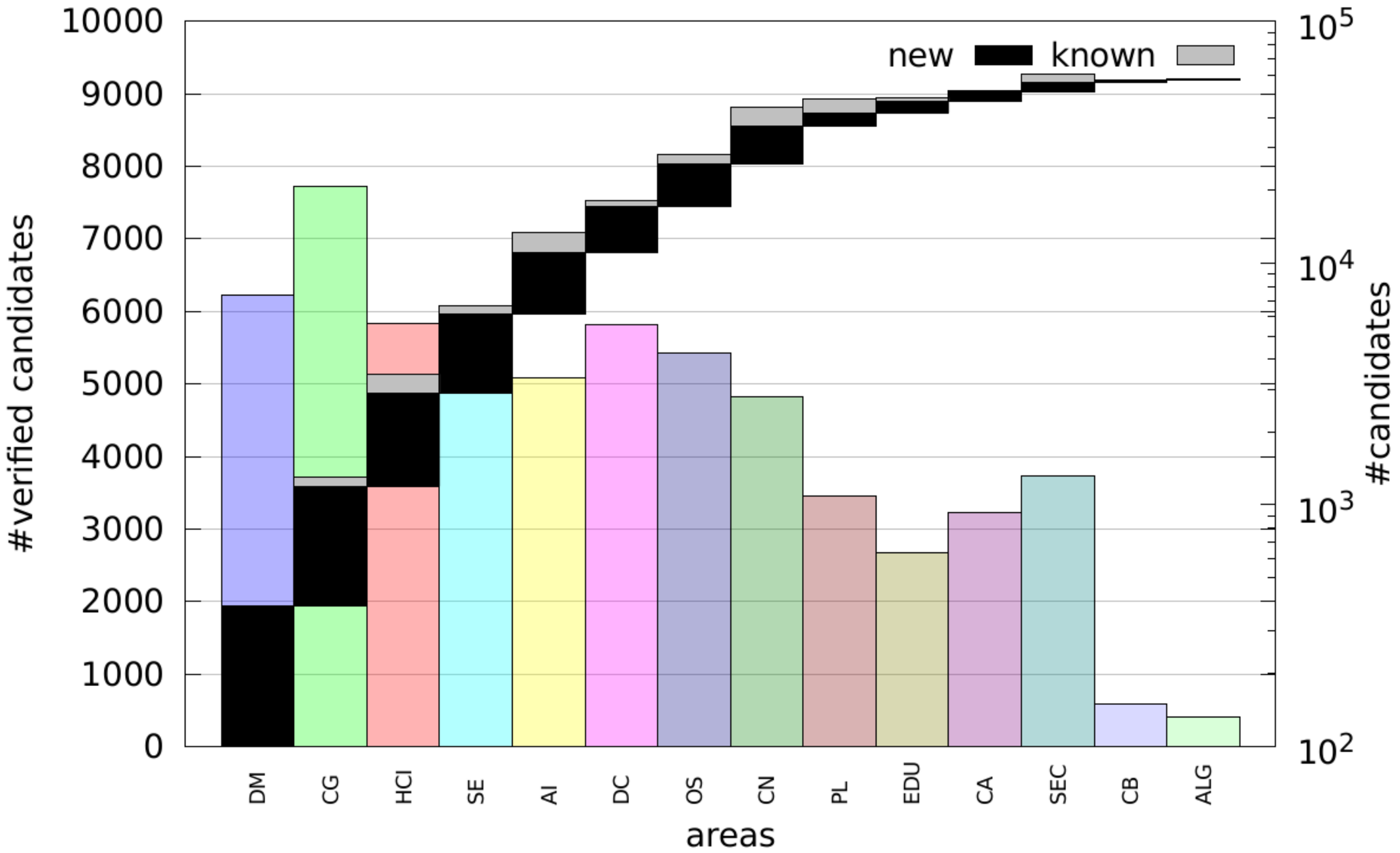
- Motivation

- Approach

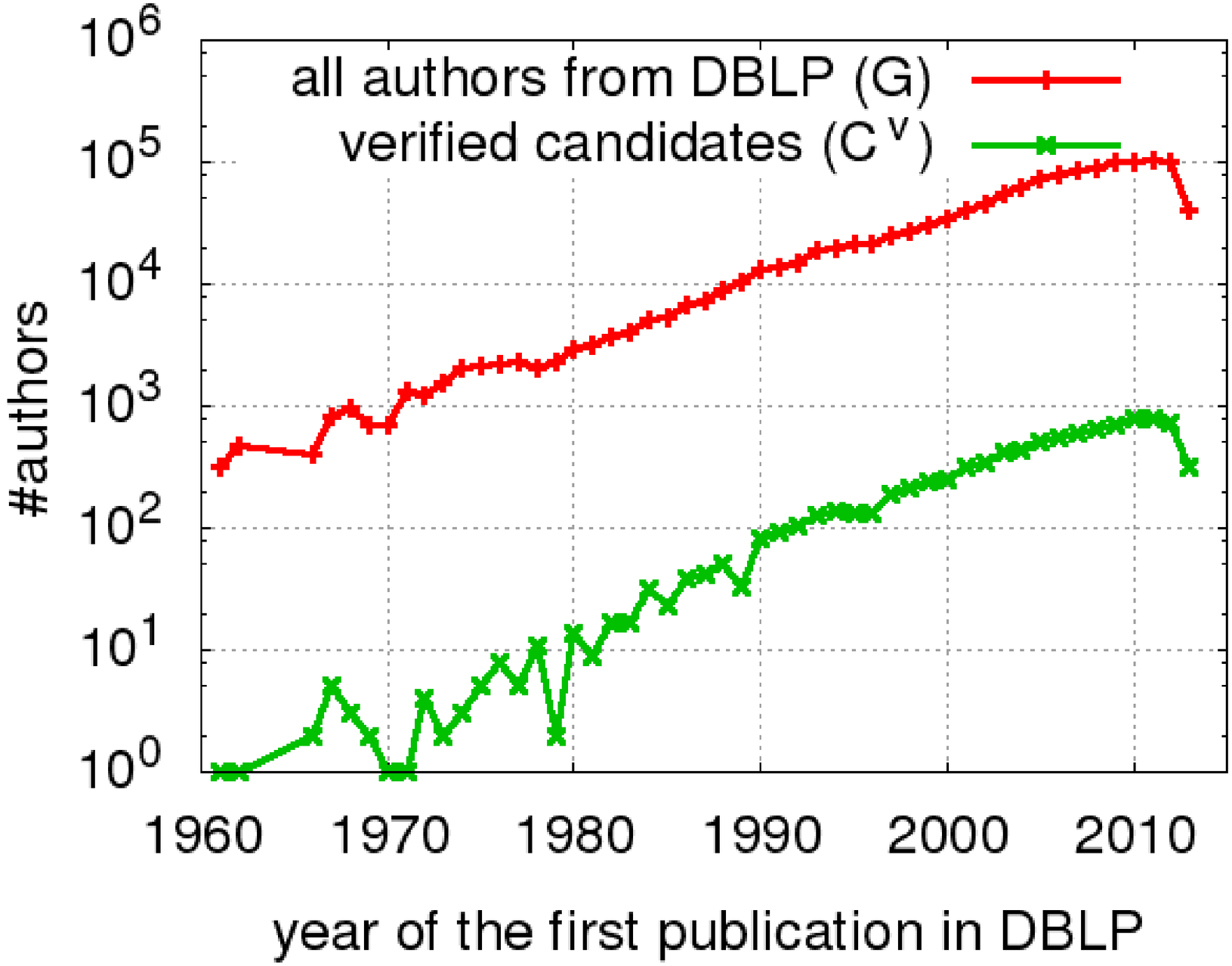


- **Results**

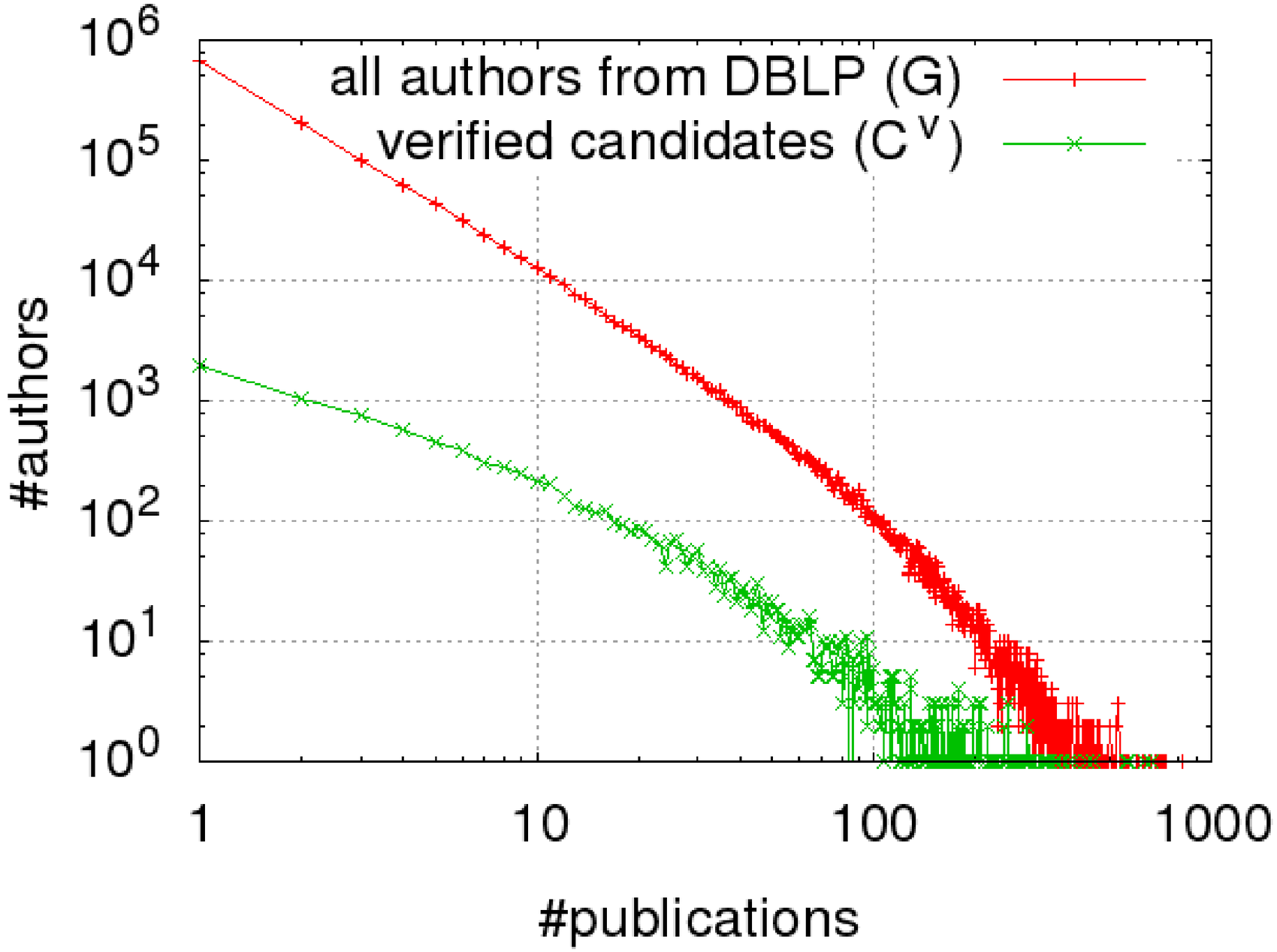
Which areas of computer science?



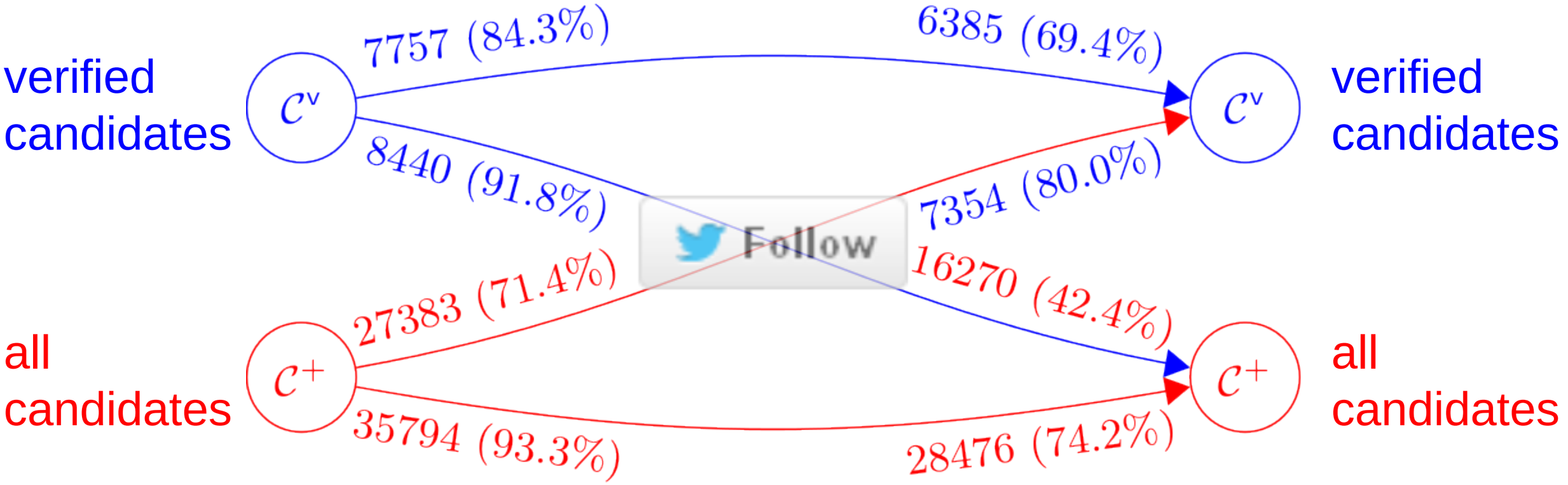
Are researchers on Twitter younger?



Are they more productive?



How are they connected with each other?

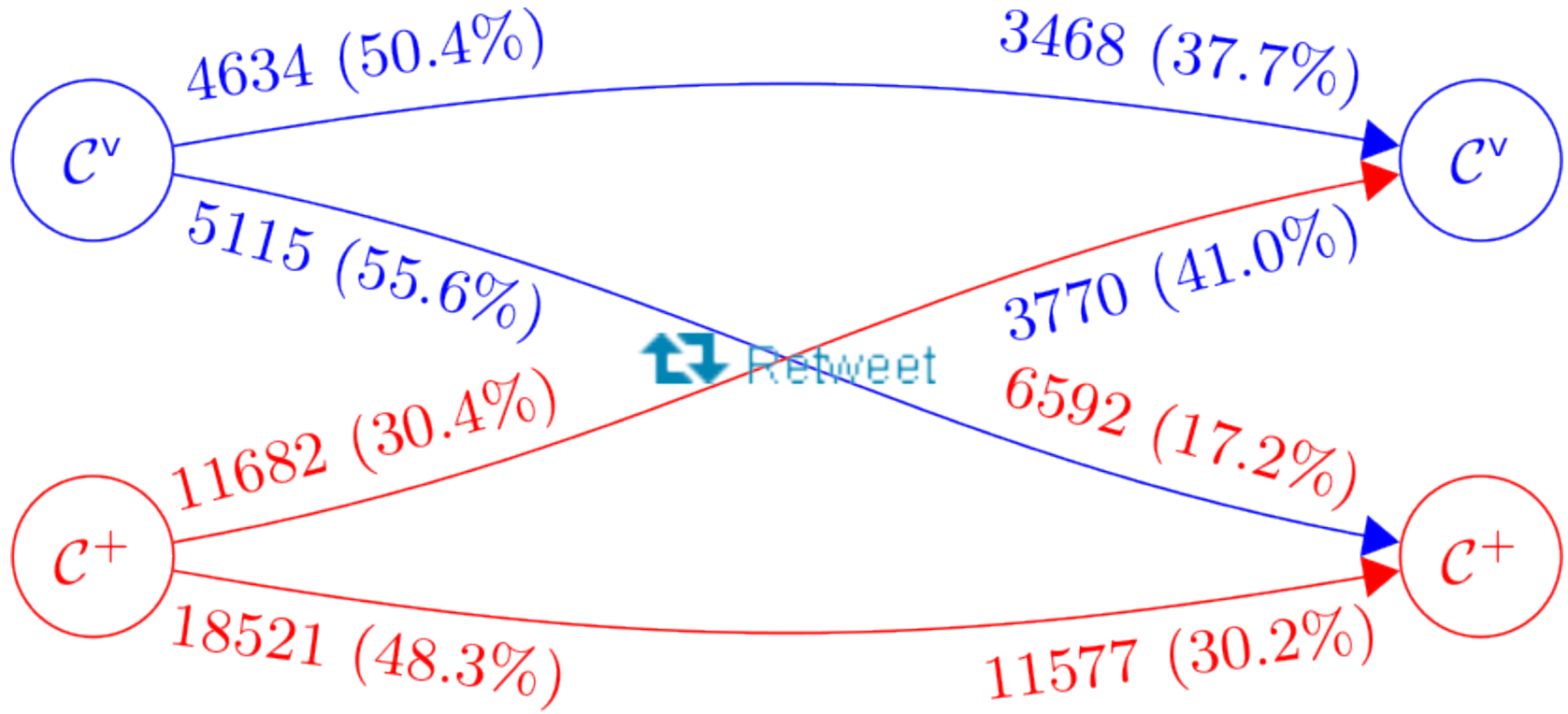


(a) Who follows whom?

- in general, the order of activity is follow, mention, retweet

How are they connected with each other?

verified candidates



verified candidates

all candidates

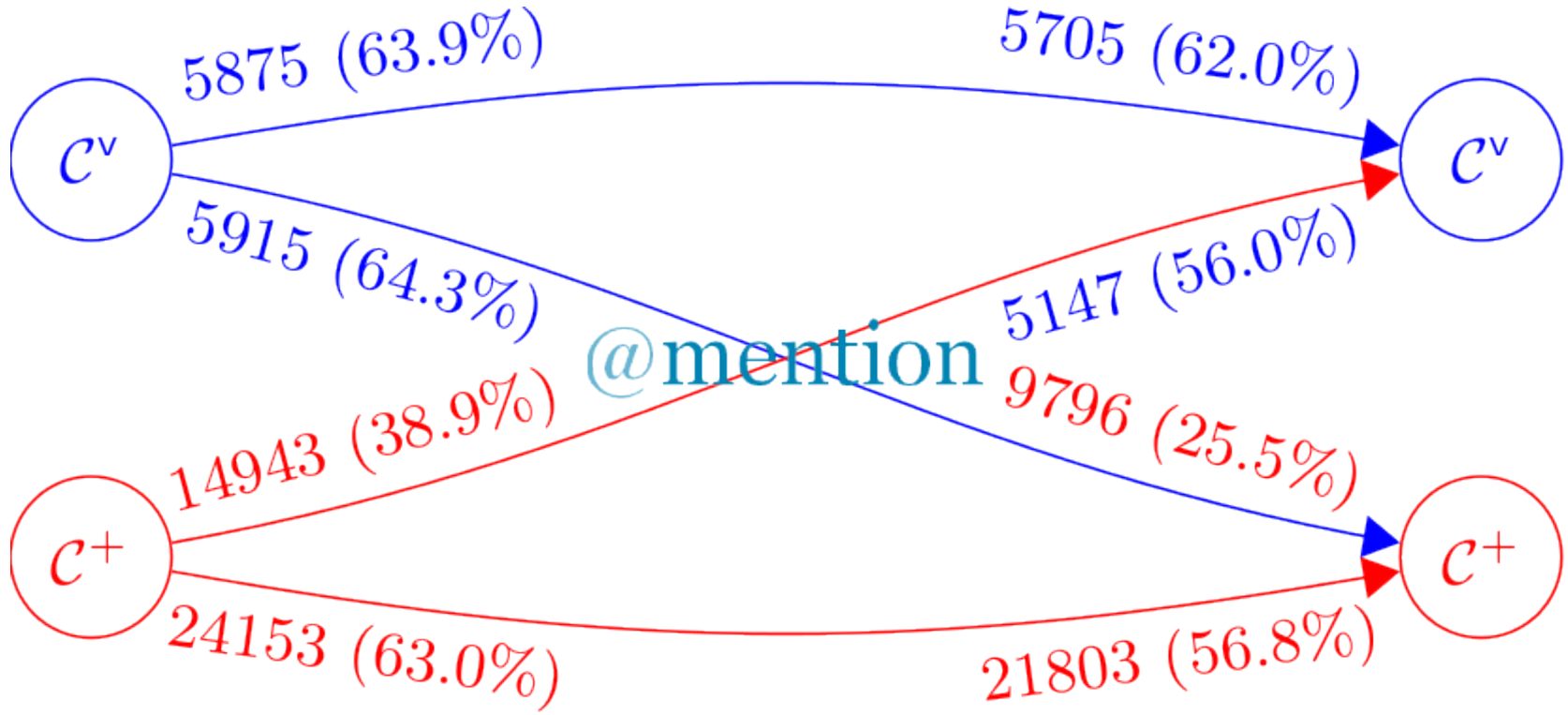
all candidates

(b) Who retweets whom?

- in general, the order of activity is follow, mention, retweet

How are they connected with each other?

verified candidates



verified candidates

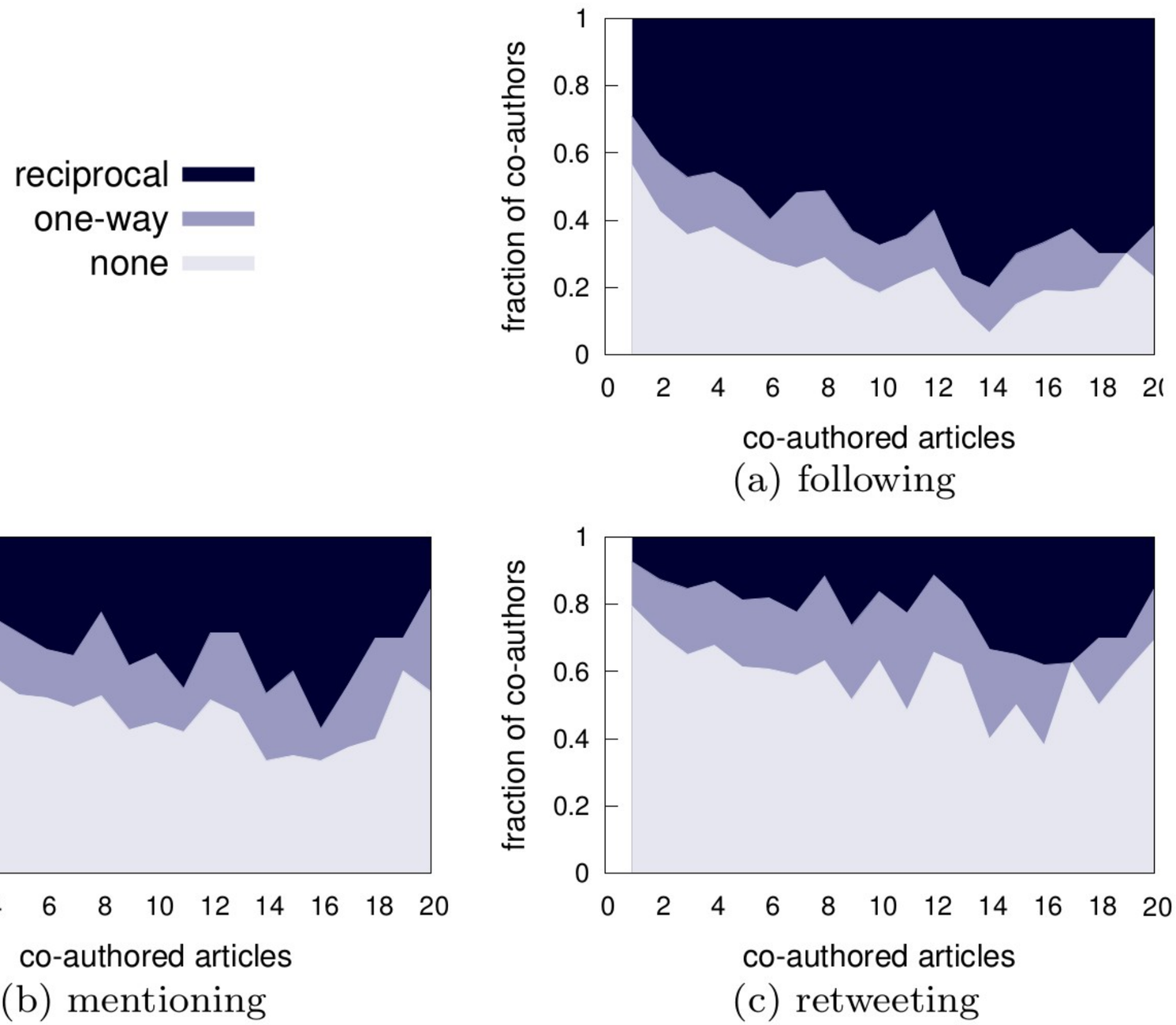
all candidates

all candidates

(c) Who mentions whom?

- in general, the order of activity is follow, mention, retweet

How does closer scientific collaboration affect interaction on Twitter?



(a) following

(b) mentioning

(c) retweeting

Who are the most influential researchers?

verified candidates

screen name	name	ranking by r	f	m	p
@timoreilly	Tim O'Reilly	1	2	1	16
@billgates	Bill Gates	2	1	2	1
@hmason	Hilary Mason	3	9	3	2
@zephoria	Danah Boyd	4	7	6	24
@csoghoian	Christopher Soghoian	5	51	12	5
@doctorow	Cory Doctorow	6	16	4	2
@ioerror	Jacob Appelbaum	7	30	7	5
@mattmight	Matthew Might	8	47	16	34
@kentbeck	Kent Beck	9	18	17	35
@mattcutts	Matt Cutts	10	15	9	2
@timberners_lee	Tim Berners-Lee	11	3	5	35
@codepo8	Christian Heilmann	12	87	14	1
@mattblaze	Matt Blaze	13	60	25	72
@digiphile	Alex Howard	14	42	13	1

retweet follow mention #publications

Who are the most influential researchers?

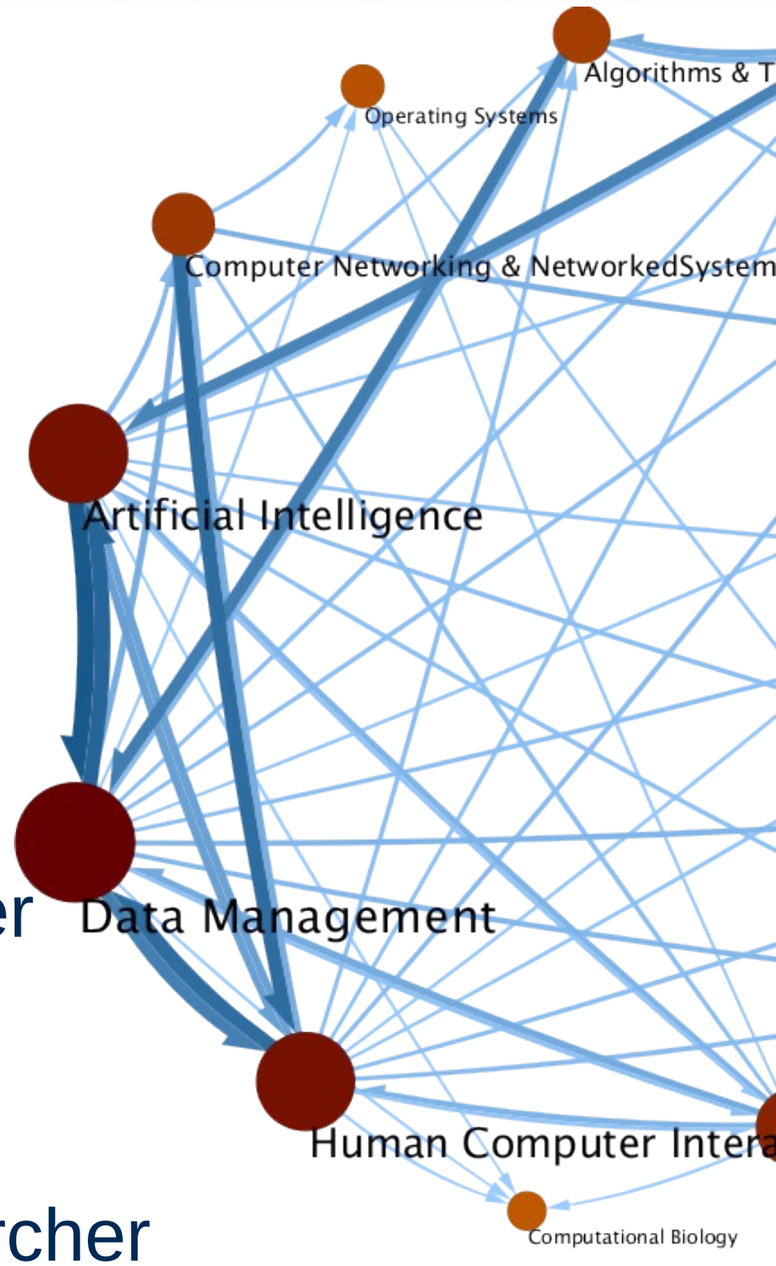
top 200 influential researchers

other researchers

terms from the user profiles

Outlook

- improve matching accuracy
- analyze topics & interests of users
- social network analysis
- transfer to other disciplines
- build a web directory of researchers on Twitter

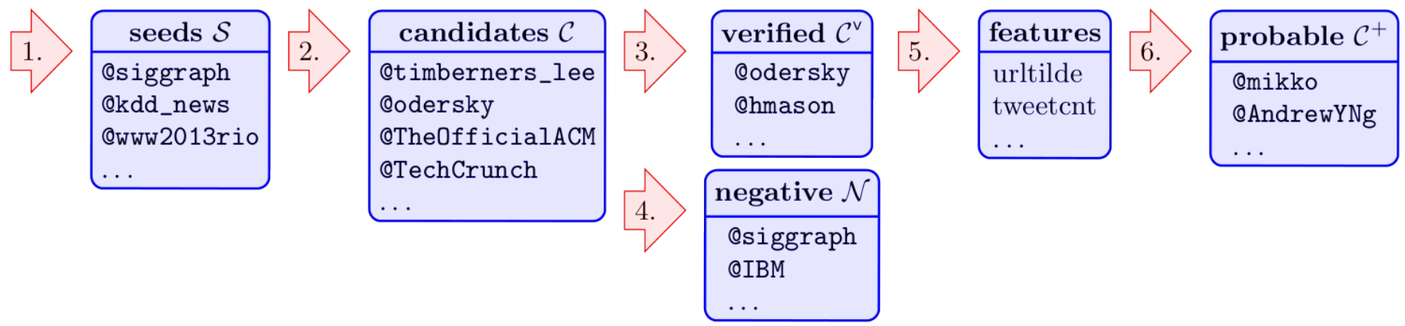


- **dataset**: https://github.com/L3S/twitter-researcher
- **paper**: Hadgu, A.T. & Jäschke, R. (2014), Identifying and Analyzing Researchers on Twitter. *Proceedings of the Web Science Conference*, New York, NY, USA: ACM.

Thanks for your patience! Questions?

- Motivation

- Approach



- Results