Where's @wally? A Classification Approach to Geolocating Twitter Users

Dominic Rout, Daniel Preoțiuc-Pietro, Kalina Bontcheva, Trevor Cohn, The University of Sheffield

The University Of Sheffield.

Finding social network users

Discovering automatically where in the physical world a Twitter lives/is posting from.

Social Science Where the location of a user can affect their experiences, upbringing, health.

Linguistics Regional vocabulary, dialog and grammar.

Advertising Serving ads about products and services that are close to the reader.

Security Detecting spammers and fake accounts.

Conferences Demographics of accepted papers.

Finding social network users

Discovering automatically where in the physical world a Twitter lives/is posting from.

Social Science Where the location of a user can affect their experiences, upbringing, health.

Linguistics Regional vocabulary, dialog and grammar.

Advertising Serving ads about products and services that are close to the reader.

Security Detecting spammers and fake accounts.

Conferences Demographics of accepted papers.

Finding social network users

Discovering automatically where in the physical world a Twitter lives/is posting from.

Social Science Where the location of a user can affect their experiences, upbringing, health.

Linguistics Regional vocabulary, dialog and grammar.

Advertising Serving ads about products and services that are close to the reader.

Security Detecting spammers and fake accounts.

Conferences Demographics of accepted papers.

Finding social network users

Discovering automatically where in the physical world a Twitter lives/is posting from.

Social Science Where the location of a user can affect their experiences, upbringing, health.

Linguistics Regional vocabulary, dialog and grammar.

Advertising Serving ads about products and services that are close to the reader.

Security Detecting spammers and fake accounts.

Conferences Demographics of accepted papers.

Finding social network users

Discovering automatically where in the physical world a Twitter lives/is posting from.

Social Science Where the location of a user can affect their experiences, upbringing, health.

Linguistics Regional vocabulary, dialog and grammar.

Advertising Serving ads about products and services that are close to the reader.

Security Detecting spammers and fake accounts.

Conferences Demographics of accepted papers.

Contributions

We approach the task of geolocation of social network users, making the following key contributions.

- We formulate user geolocation as a classification problem, with a limited search space.
- We demonstrate a powerful, simplifying assumption for the geolocation task.
- We present a system for user geolocation that is easier to apply than existing approaches.
- We formulate a new data set for geolocation in the U.K. available online (anonymised).

Focus here on Twitter - actual problem is more general!

The location of a Twitter post

GPS Co-ordinates

- When posting from a mobile phone.
- Attached to a single tweet.
- Can be searched over.
- Not necessarily where you live.

Profile fields

- Author of post creates a short biography, or ‘profile’.
- Can include a ‘location’ - which is unrestricted free-text.

Why geolocate automatically?

Only a small portion of tweets have GPS co-ordinates attached. Profile fields are useful, but many leave them blank, or make up locations:

Locations of users of Hashtag #yolo

Chicago, IL
Jakarta, Indonesia
In the moment
la la land
Madchester
ObviouslyNotSingapore

#YOLO

Abbreviation for: you only live once

The idiots's excuse for something stupid that they did.

"Hey i heard u got that girl pregnant"

"Ya man but hey YOLO"

From `www.urbandictionary.com`

Related Work

Cheng Using textual content of Tweets to locate users.

Mahmud Ensemble of location classifiers using textual features.

Eisenstein Latent variable model associating vocabulary with locations.

Backstrom Finding users on Facebook using social network.

We propose a new social network-based approach relying upon lightweight features and using a reduced search space.

Related Work

Cheng Using textual content of Tweets to locate users.

Mahmud Ensemble of location classifiers using textual features.

Eisenstein Latent variable model associating vocabulary with locations.

Backstrom Finding users on Facebook using social network.

We propose a new social network-based approach relying upon lightweight features and using a reduced search space.

Related Work

Cheng Using textual content of Tweets to locate users.

Mahmud Ensemble of location classifiers using textual features.

Eisenstein Latent variable model associating vocabulary with locations.

Backstrom Finding users on Facebook using social network.

We propose a new social network-based approach relying upon lightweight features and using a reduced search space.

Related Work

Cheng Using textual content of Tweets to locate users.

Mahmud Ensemble of location classifiers using textual features.

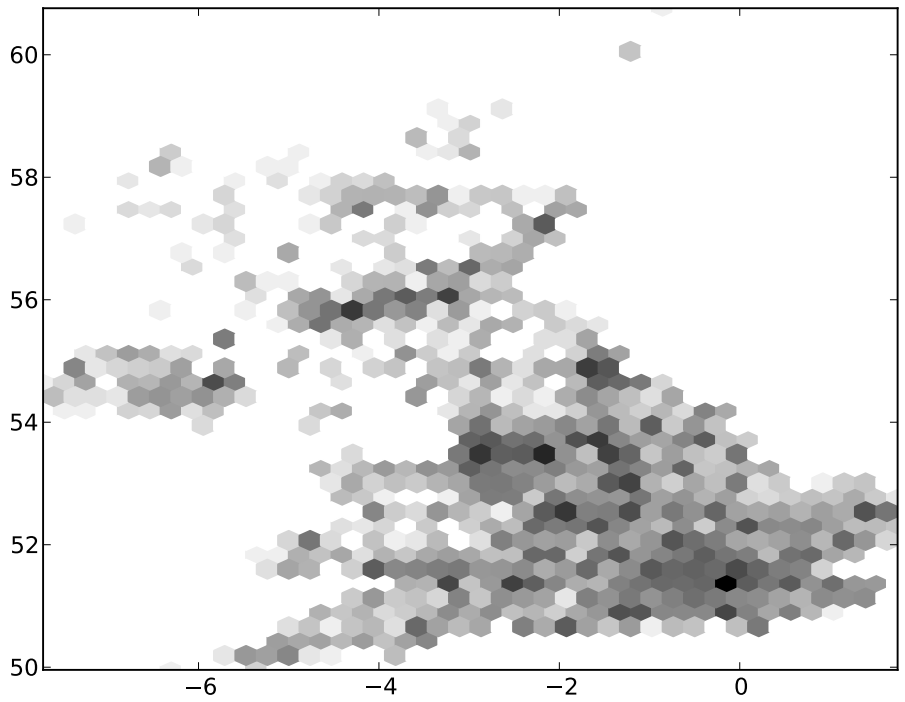
Eisenstein Latent variable model associating vocabulary with locations.

Backstrom Finding users on Facebook using social network.

We propose a new social network-based approach relying upon lightweight features and using a reduced search space.

UK Twitterers

- We have built a data set of Twitter members from the UK
- This forms our evaluation data set.
- All users have known locations
- Hide location knowledge for evaluation.



The University Of Sheffield.

Understanding profile locations

We use a simple, high-precision method to parse location names and resolve against an ontology of UK towns.

1. Country suffixes are tested against a list of synonyms for the UK.
2. Country names are then truncated from the original string.
3. Cities are matched against names from DBpedia.
4. In case of conflict, more populous cities (according to the ontology) are assumed.

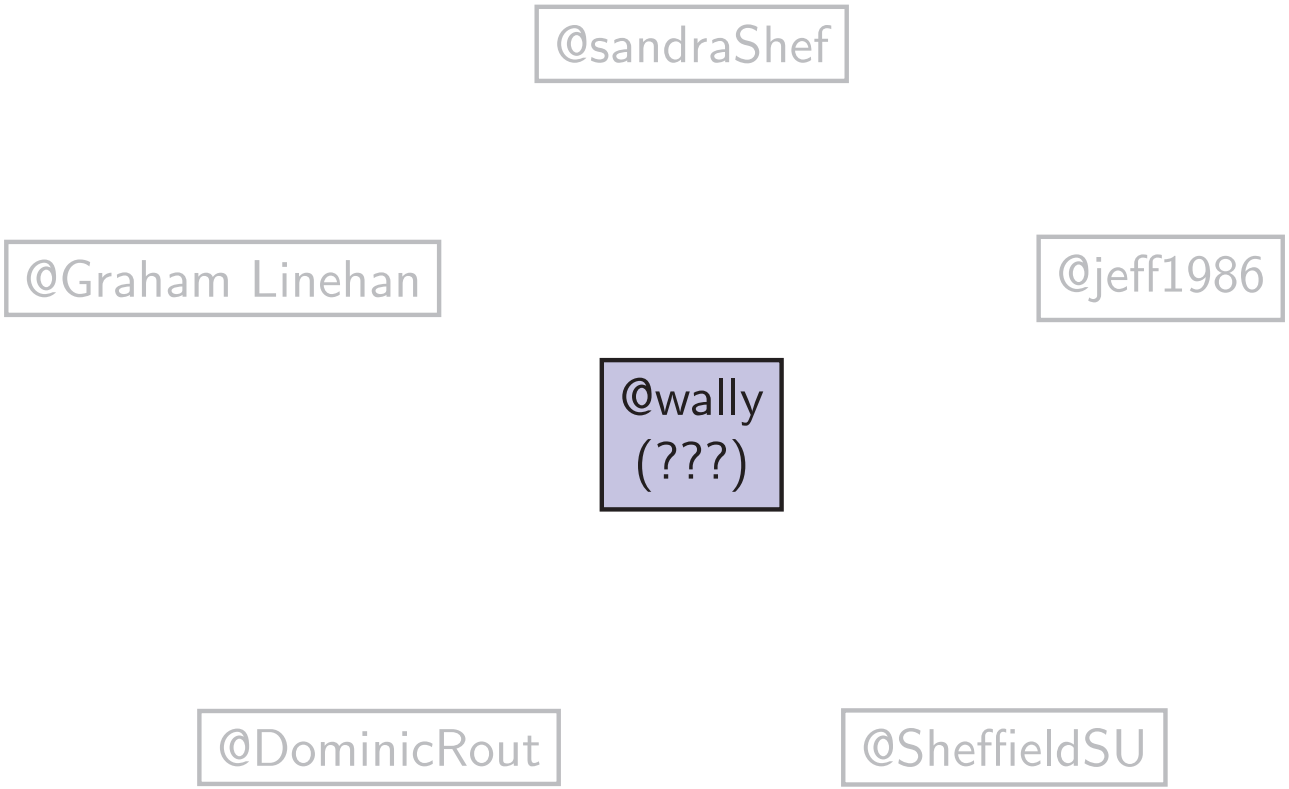
This approach works in the UK, where city and town names are largely unambiguous. Hecht et al (2011) study location fields more intimitely.

The social graph

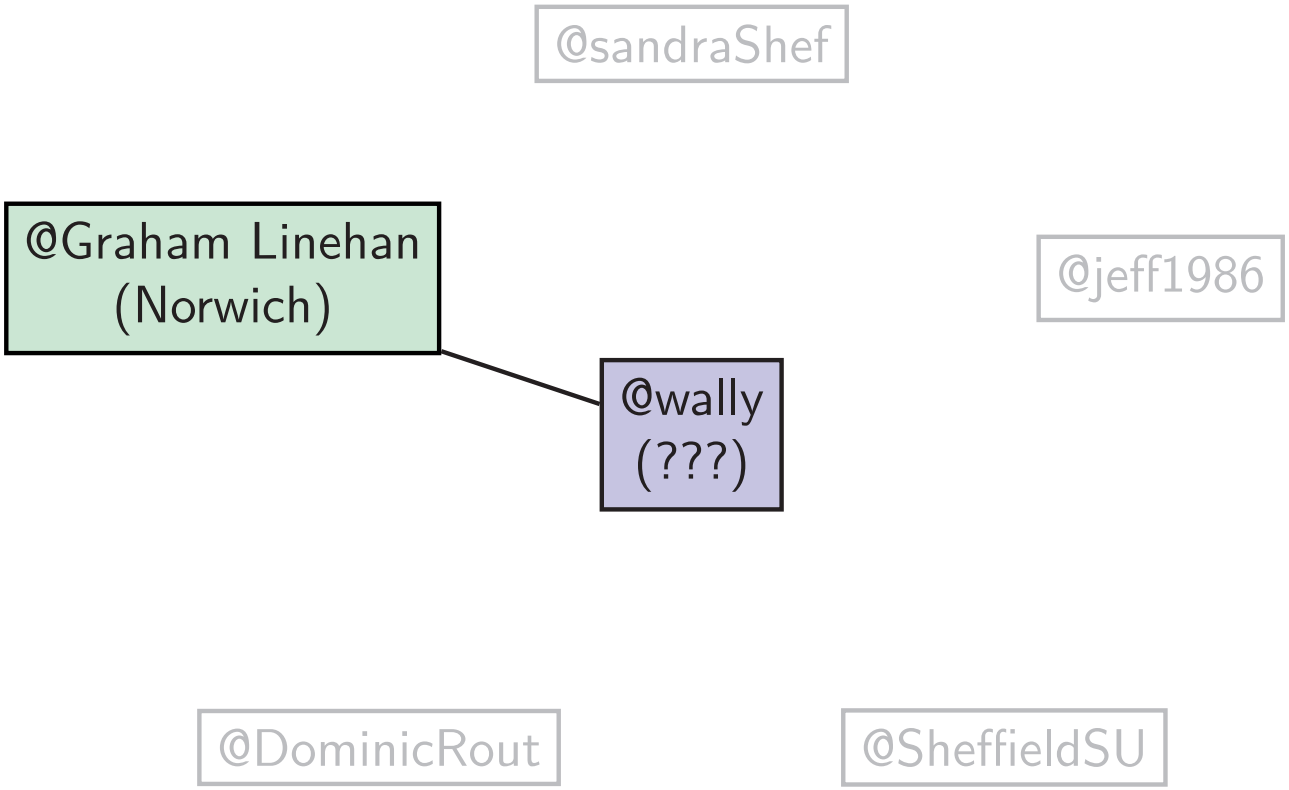
For our method, we discover the users, their locations and the connections between them.

- Use Twitter API to download part of the social graph.
- The graph contains only users with easily resolved profile location strings.
  - $206,200$ users.
  - Mean in/out degree of 48.75
  - Median indegree (followers in our graph): 11
  - Median outdegree (followees in our graph): 27
  - Reciprocity: 37%

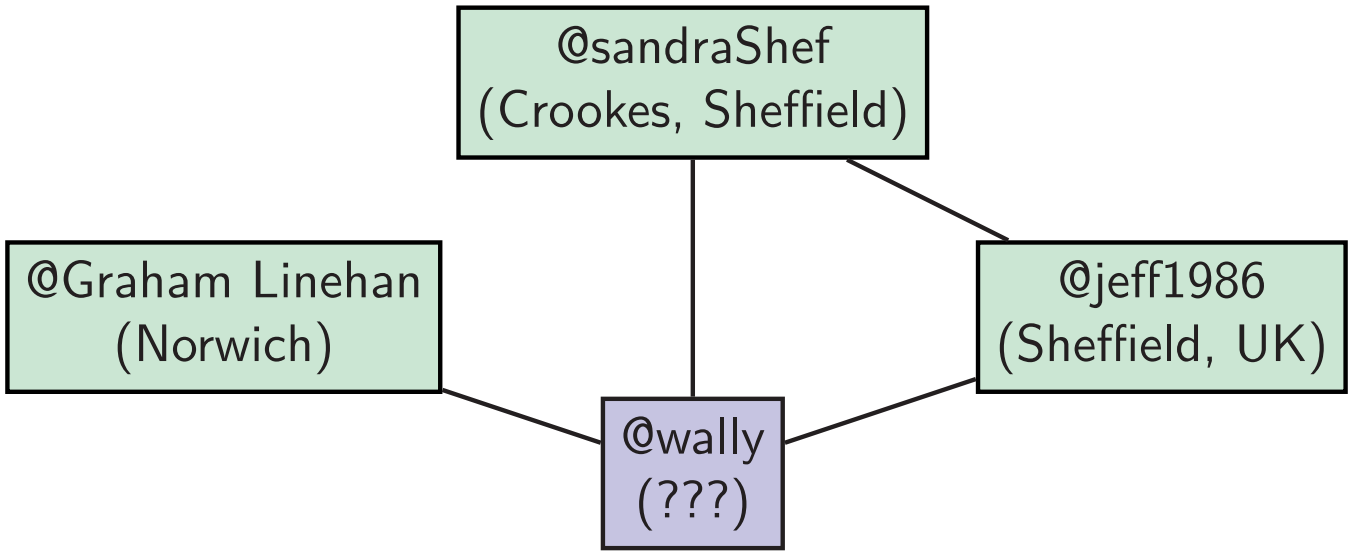
Where's @wally ?



Where's @wally ?



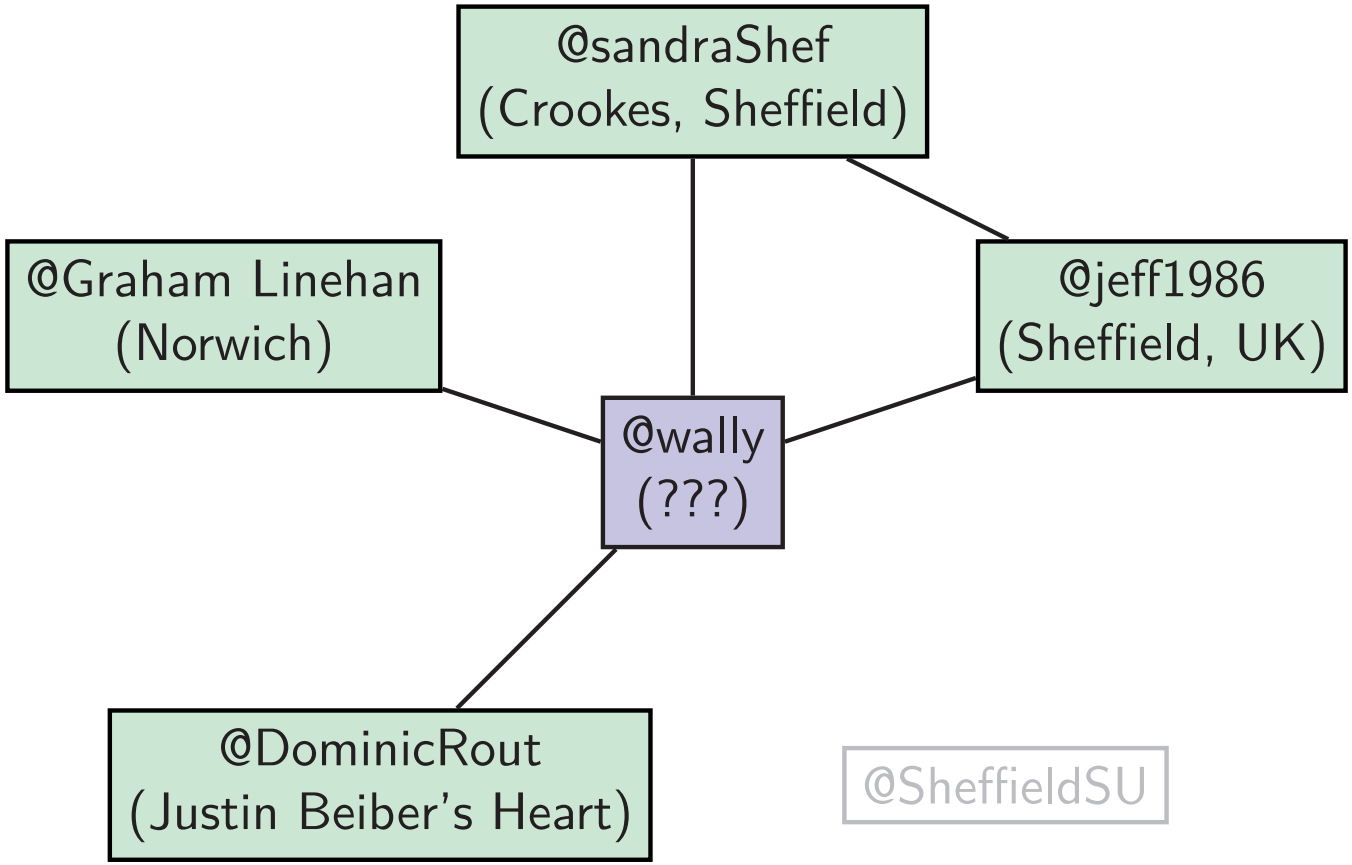
Where's @wally ?



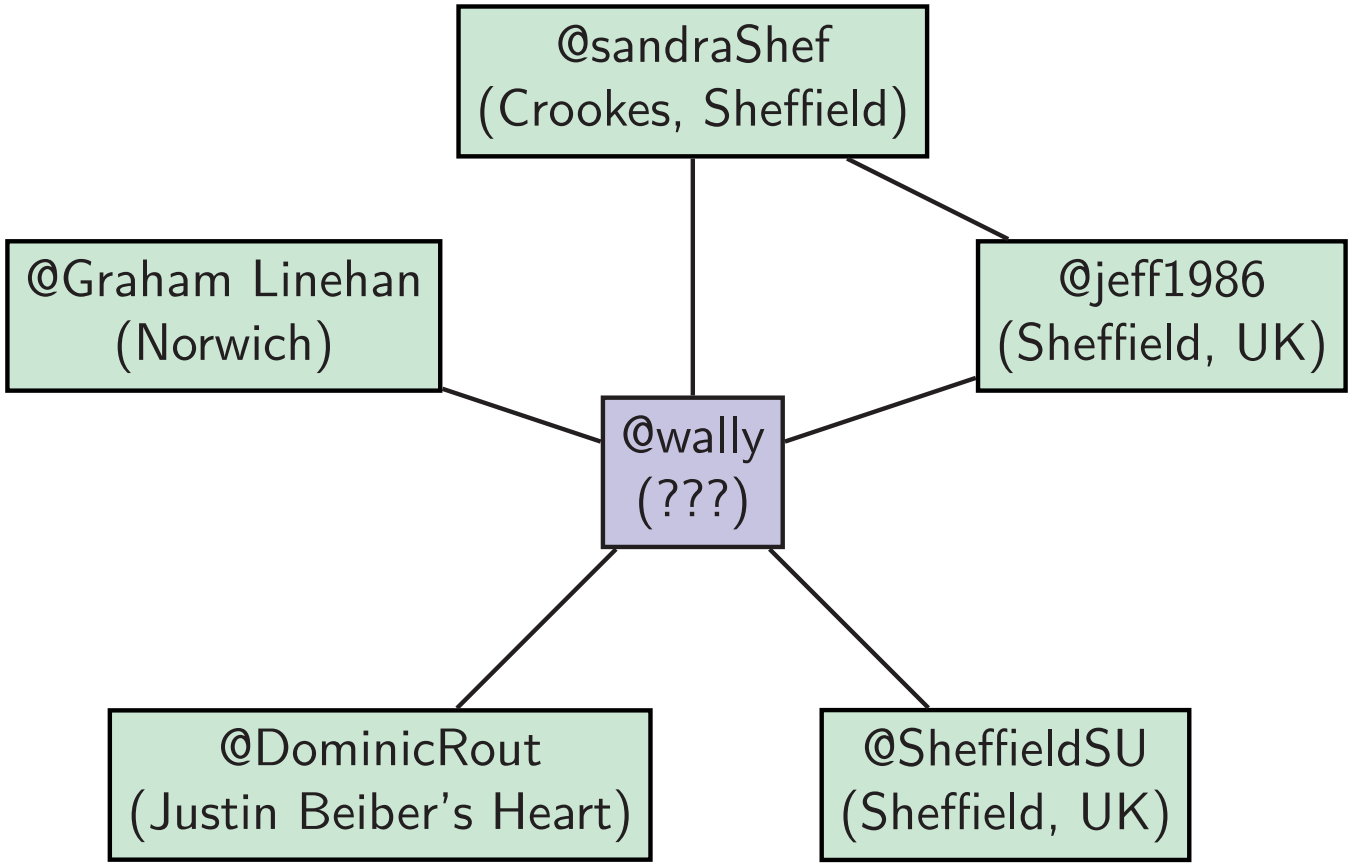
@DominicRout

@SheffieldSU

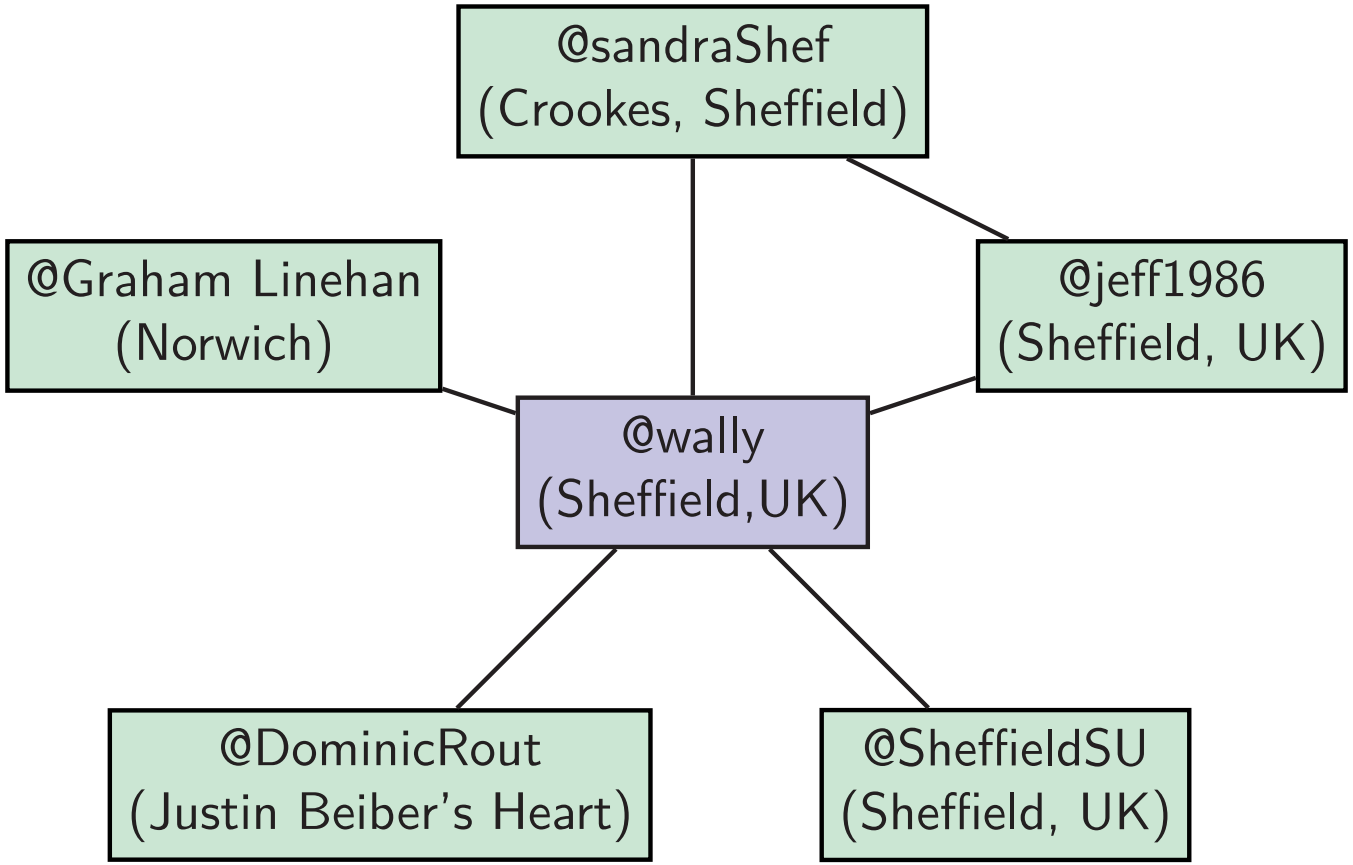
Where's @wally ?



Where's @wally ?

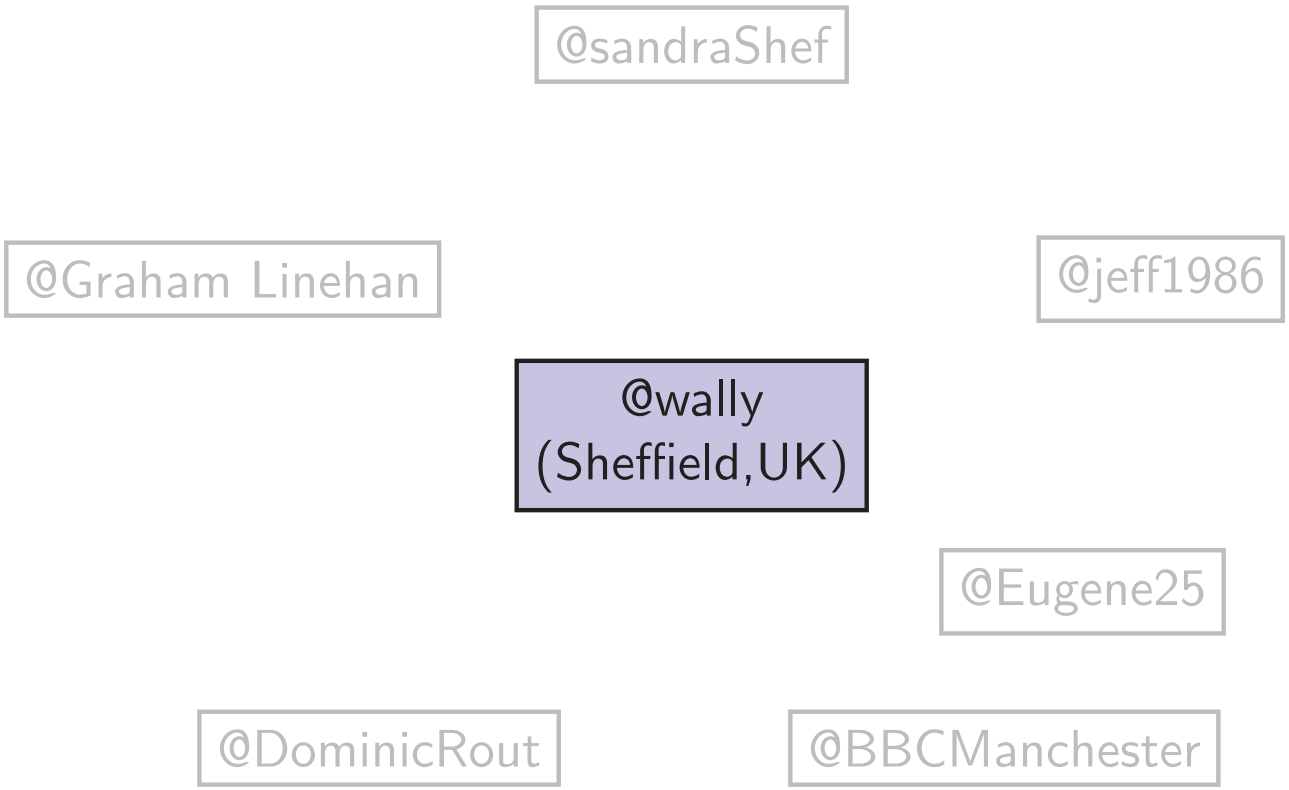


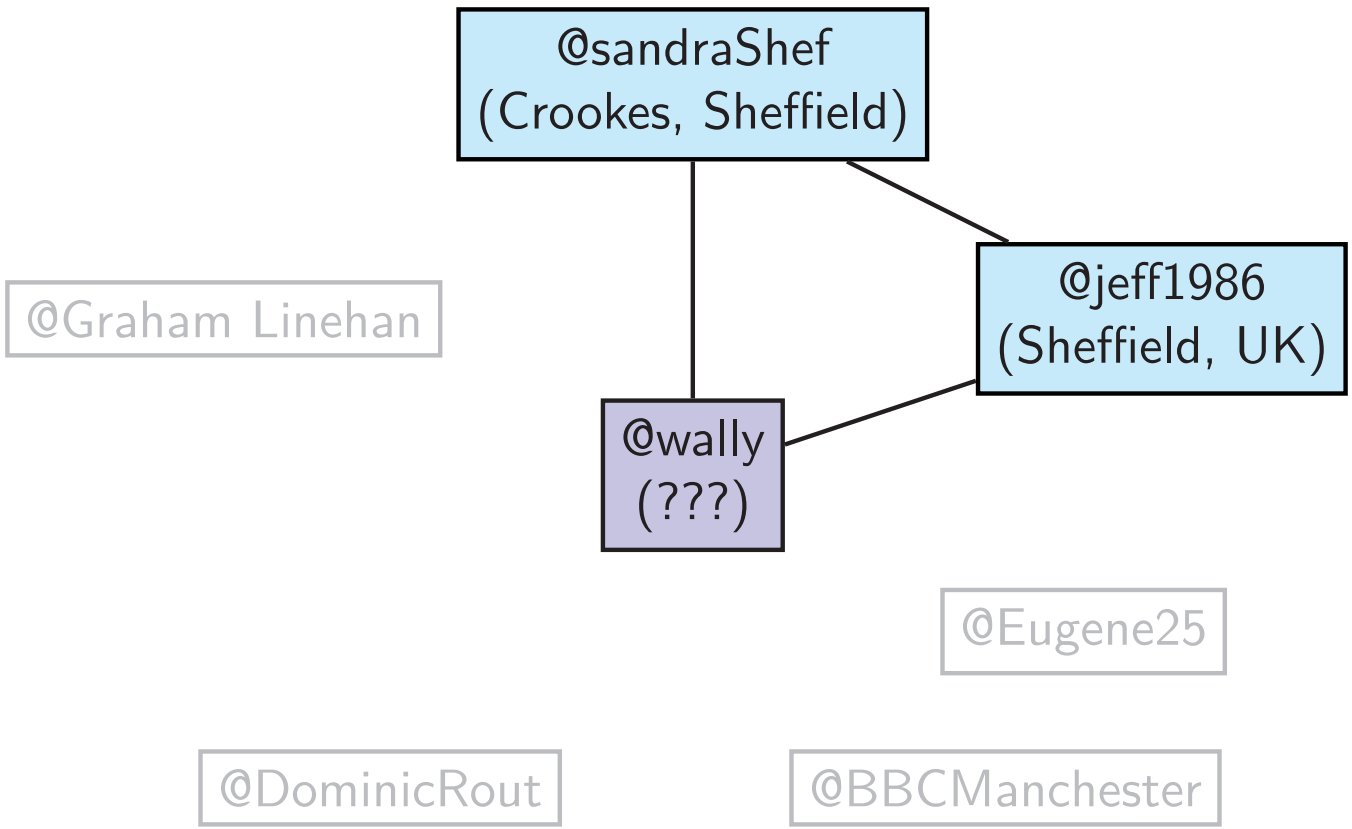
Where's @wally ?

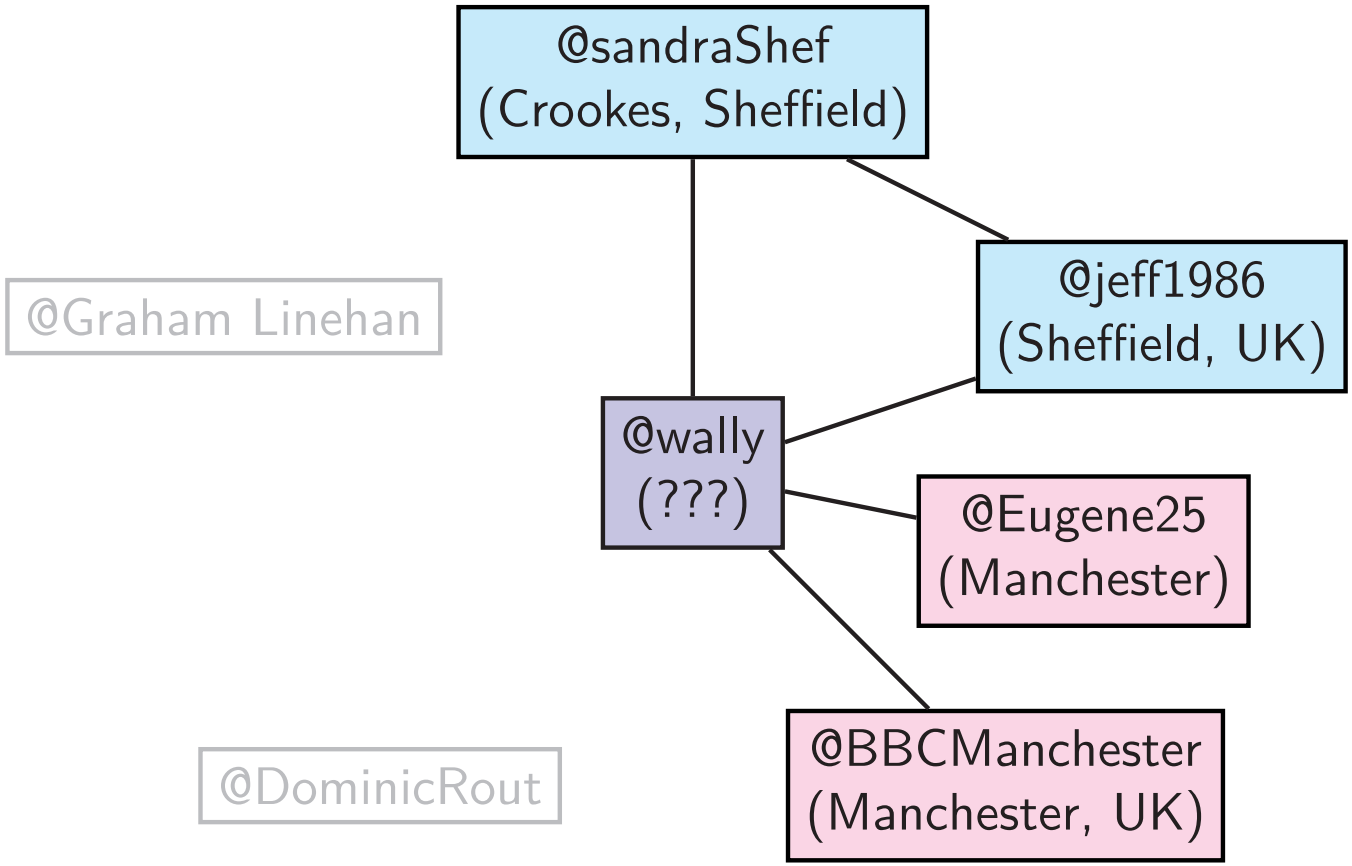


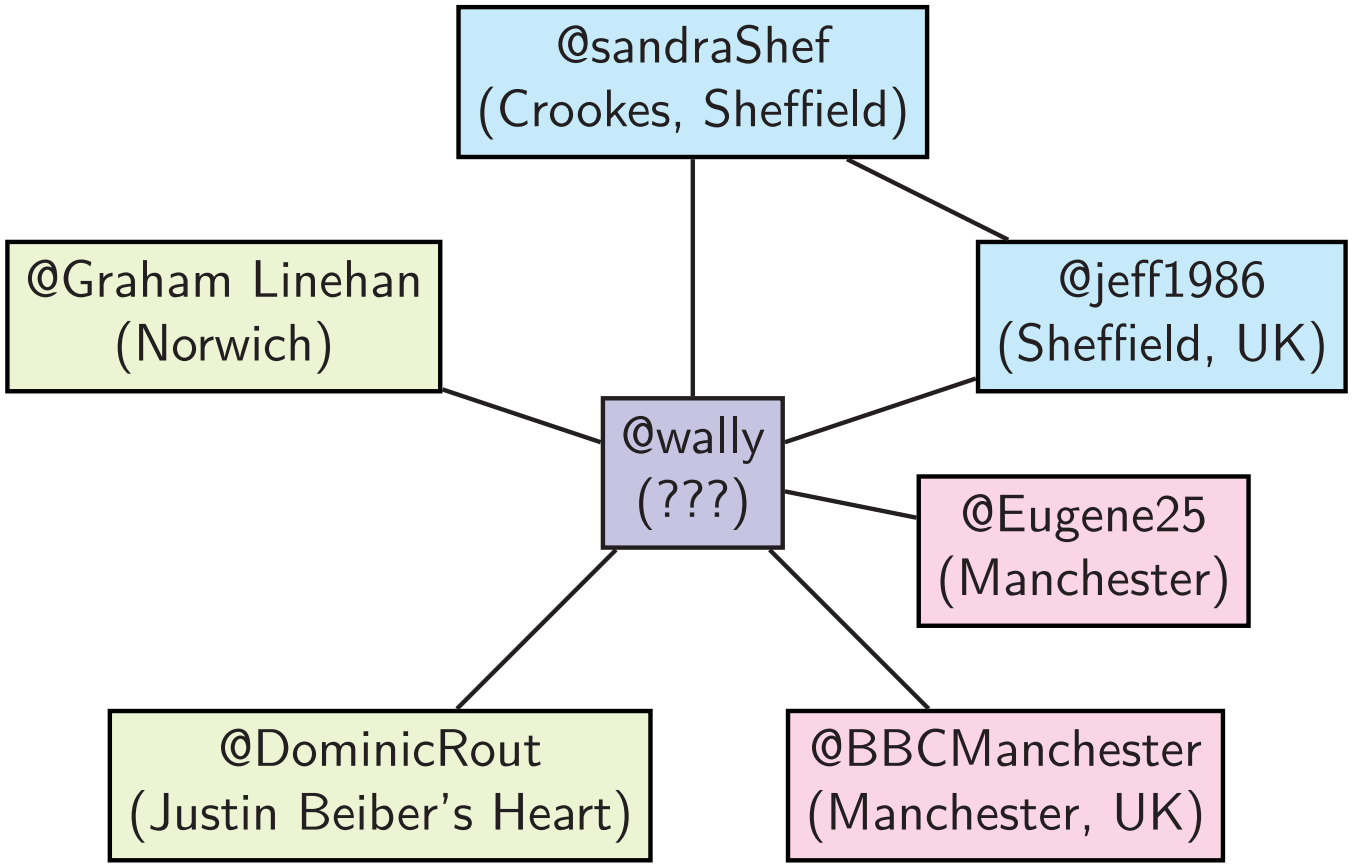
Where's @wally ?

- This user's friends are mostly in Sheffield
- So we might assume that they live in Sheffield
- This is a kind of homophily.
- Just counting users works $39.49\% \pm 0.39\%$ of the time (on our data set)


@sandraShef
@Graham_Linehan
@jeff1986
@wally
(Sheffield,UK)
@Eugene25
@DominicRout
@BBCManchester


@sandraShef
(Crookes, Sheffield)
@jeff1986
(Sheffield, UK)
@wally
(???)
@Graham Linehan
@Eugene25
@DominicRout
@BBCManchester


@sandraShef
(Crookes, Sheffield)
@jeff1986
(Sheffield, UK)
@Graham Linehan
@wally
(???)
@Eugene25
(Manchester)
@BBCManchester
(Manchester, UK)
@DominicRout


@sandraShef
(Crookes, Sheffield)
@Graham Linehan
(Norwich)
@jeff1986
(Sheffield, UK)
@wally
(???)
@Eugene25
(Manchester)
@DominicRout
(Justin Beiber's Heart)
@BBCManchester
(Manchester, UK)

The Value of Friendship

- Same number of users in both locations
- Certain users might be more indicative
- Wally seems to be part of a friendship group with Sandra and Jeff
- However his friendship with BBCManchester is not part of a group - and probably not reciprocated!
- More sophisticated methods are needed to characterise these relationships.

Reducing the candidate space

We only consider locations inhabited by friends of the user.

- Reduces the candidate set from thousands of locations to just a few.
- Interested in only known locations of friends - don't attempt label propogation.
- Here we have Sheffield, London, Barnsley, York, Manchester and Walsall.

The University Of Sheffield.

Random Candidate

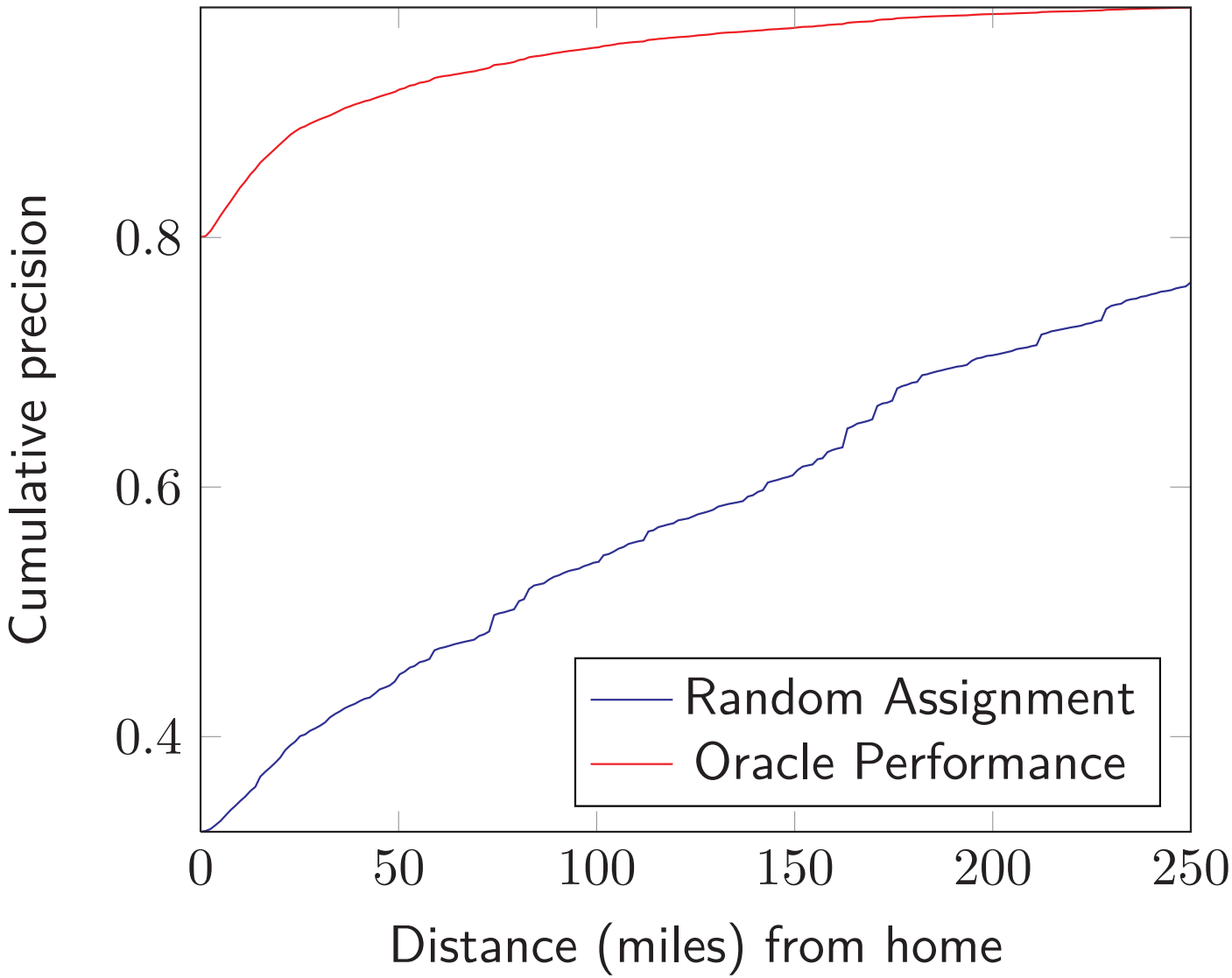
We choose some random location from the candidate set.

Oracle Performance

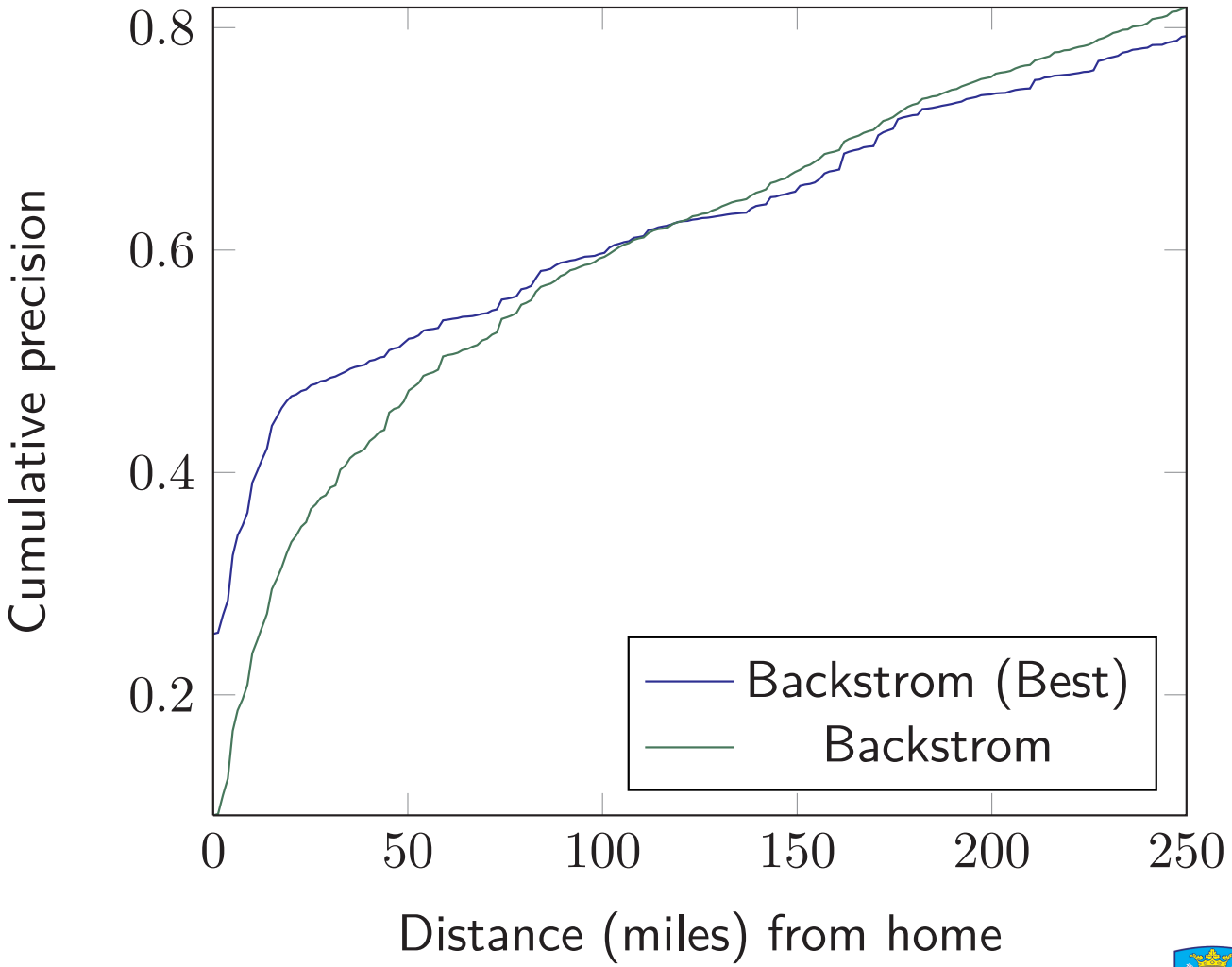
This is the upper bound on performance in our setting.

$$L'_u = \underset{c \in C_u}{\text{argmin}} \left(\text{distance}(L_u, c)\right)$$

Where E_u is the set of users following u, and L_x is the known location of user x.
Baseline is closest candidate location to true location.


Cumulative precision
0.8
0.6
0.4
0
50
100
150
200
250
Distance (miles) from home
Random Assignment
Oracle Performance

Backstrom method



Baseline methods

We implement several baseline methods, in addition to random assignment.
These baselines do not learn any classifiers.

Simple Friendship Count

Number of friends a user u has in a candidate location c.

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(\left| x \in E_u : L_x = c \right| \right)$$

Inverse City Frequency

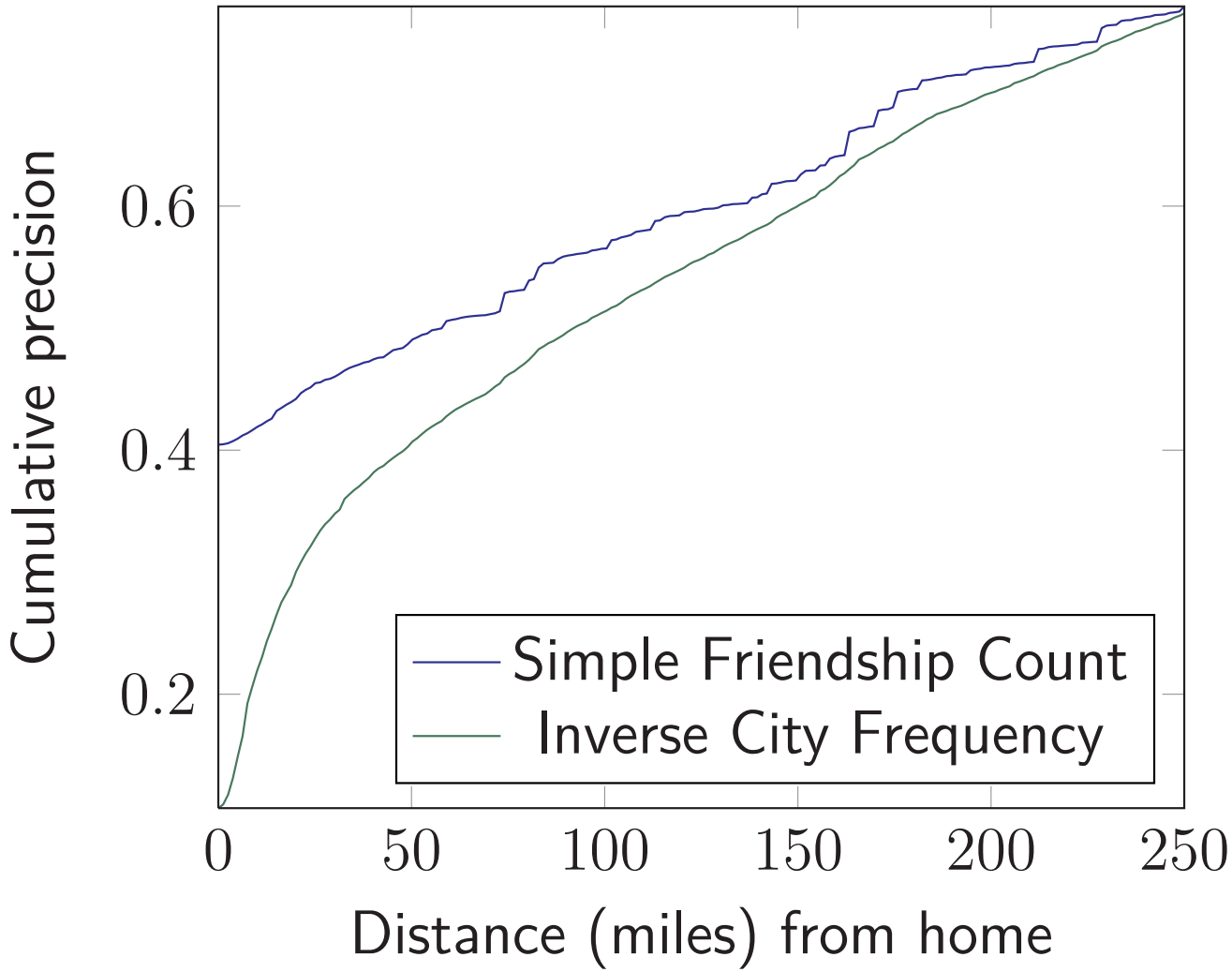
Take into account that some cities have higher population.

$$ICF_l = \frac{|U|}{|\{u \in U : L_u = l\}|}$$

And select city with highest friends $\cdot$ ICF:

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(|\{x \in E_u : L_x = c\}| \cdot ICF_c \right)$$

Population on Twitter is used to calculate ICF.



Population counts are too simple, ICF is too drastic.

Relationship Classification

Given a location l, and relationships $u \rightarrow x : L_x = l$, is l the location of user u?

- Single classifier for all locations.
- Boolean target - does location = user location?
- Features are derived from the friends of user instance
- Winner-takes-all. Explanation that gives highest score.

$$L'_u = \underset{c \in C_u}{\operatorname{argmax}} \left(S(L_x = c) \right)$$

We use SVM as provided by LIBSVM[a] for classification.

[a] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

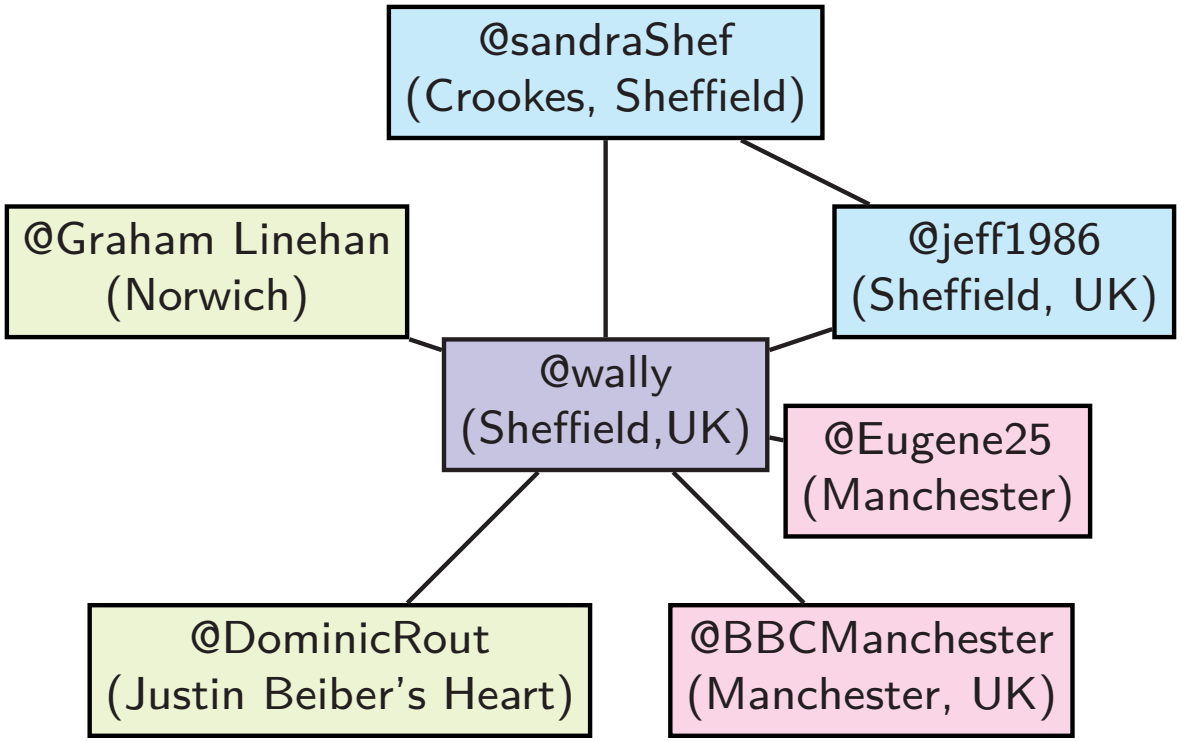
Classification Features

We implement a variety of features for classification of each location *l* according to collocation.

Simple Friendship Counts The number of friends in the location.

Simple Friendship Counts

Count number of friends in a given location.



Sheffield	2
Manchester	2
Norwich	1

Classification Features

We implement a variety of features for classification of each location *l* according to collocation.

Simple Friendship Counts The number of friends in the location.

City Population Binned value of population of the location (according to our data)

Classification Features

We implement a variety of features for classification of each location / according to collocation.

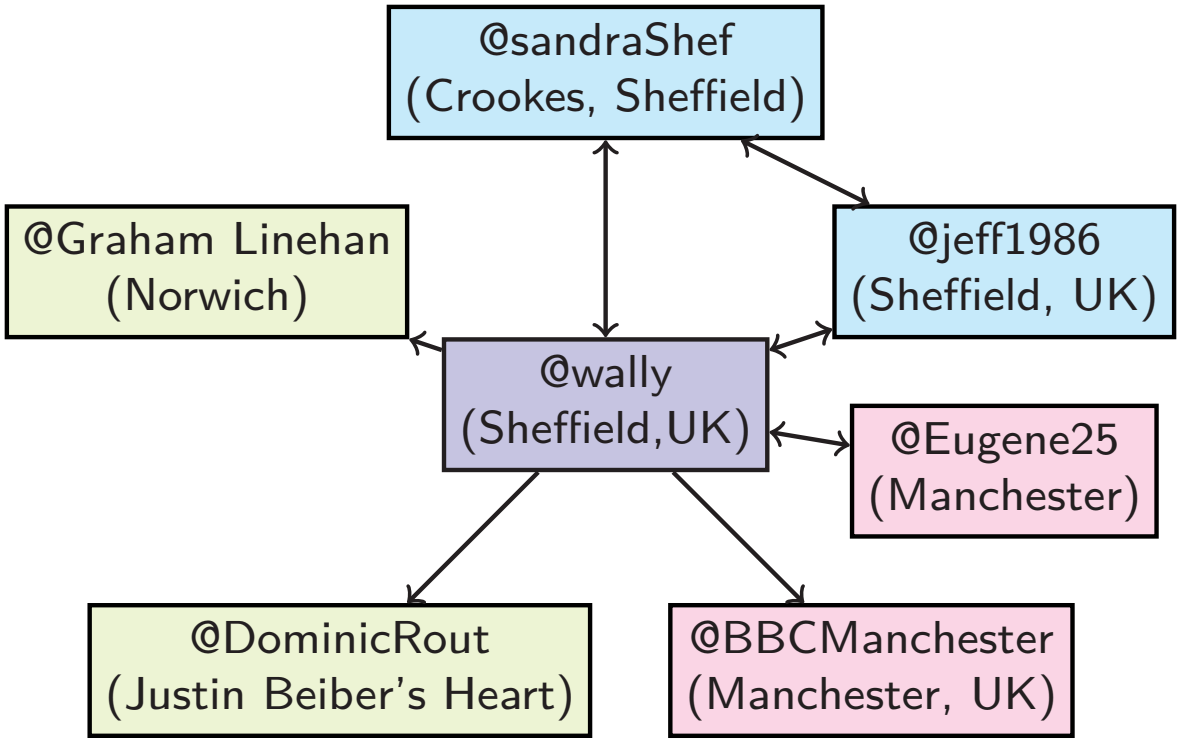
Simple Friendship Counts The number of friends in the location.

City Population Binned value of population of the location (according to our data)

Reciprocation How many relationships with people in that location are reciprocated?

Reciprocated Friendships

Count number of friends in a given location.



Sheffield	2
Manchester	1
Norwich	0

Classification Features

We implement a variety of features for classification of each location / according to collocation.

Simple Friendship Counts The number of friends in the location.

City Population Binned value of population of the location (according to our data)

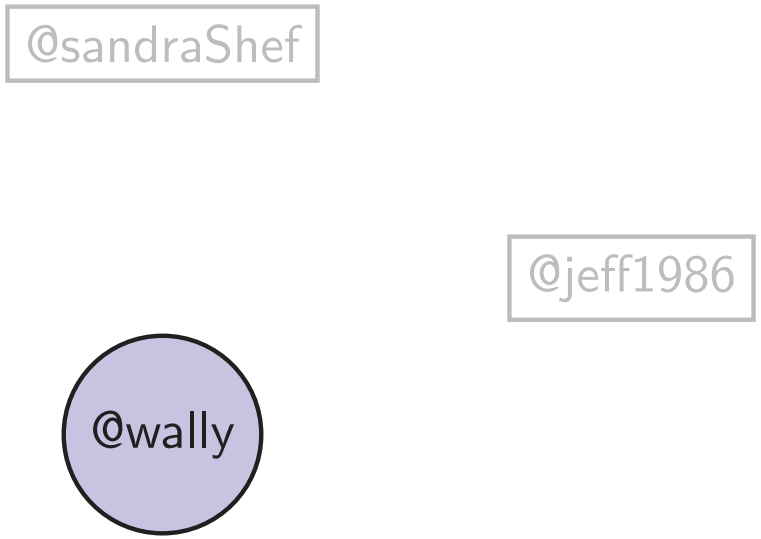
Reciprocation How many relationships with people in that location are reciprocated?

Triads How many relationships with people in that location are part of triads?

Forming Triads

People tend to befriend friends-of-friends over time.

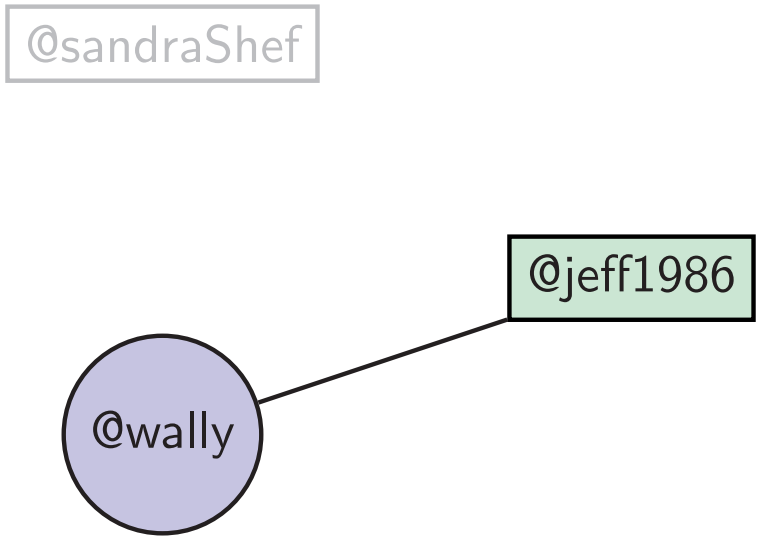
1. Wally meets Jeff
2. Jeff meets (and possibly falls for) Sandra
3. Jeff introduces Sandra to Jeff



Forming Triads

People tend to befriend friends-of-friends over time.

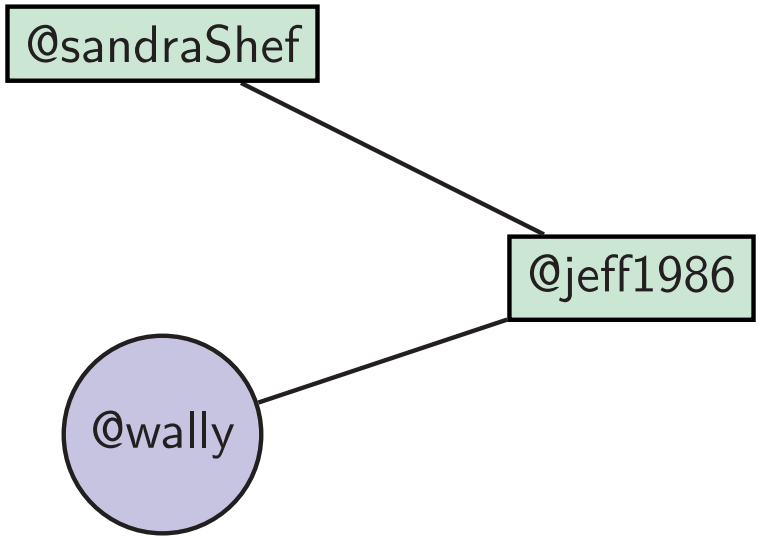
1. Wally meets Jeff
2. Jeff meets (and possibly falls for) Sandra
3. Jeff introduces Sandra to Jeff



Forming Triads

People tend to befriend friends-of-friends over time.

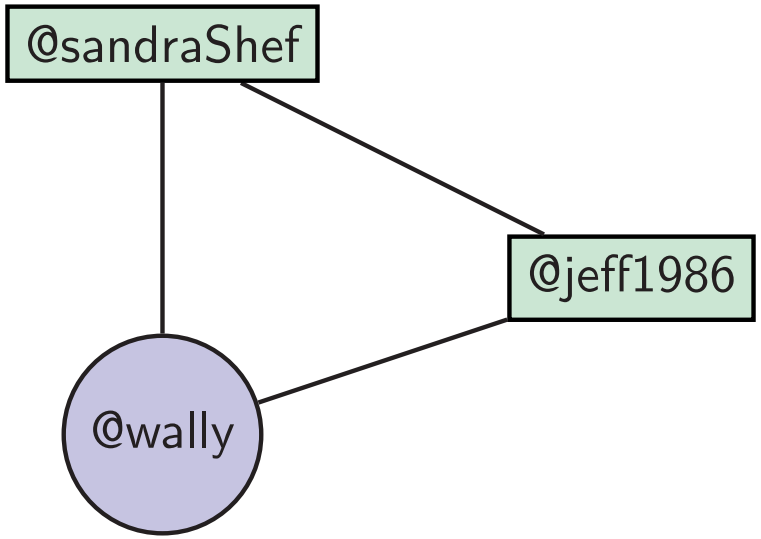
1. Wally meets Jeff
2. Jeff meets (and possibly falls for) Sandra
3. Jeff introduces Sandra to Jeff

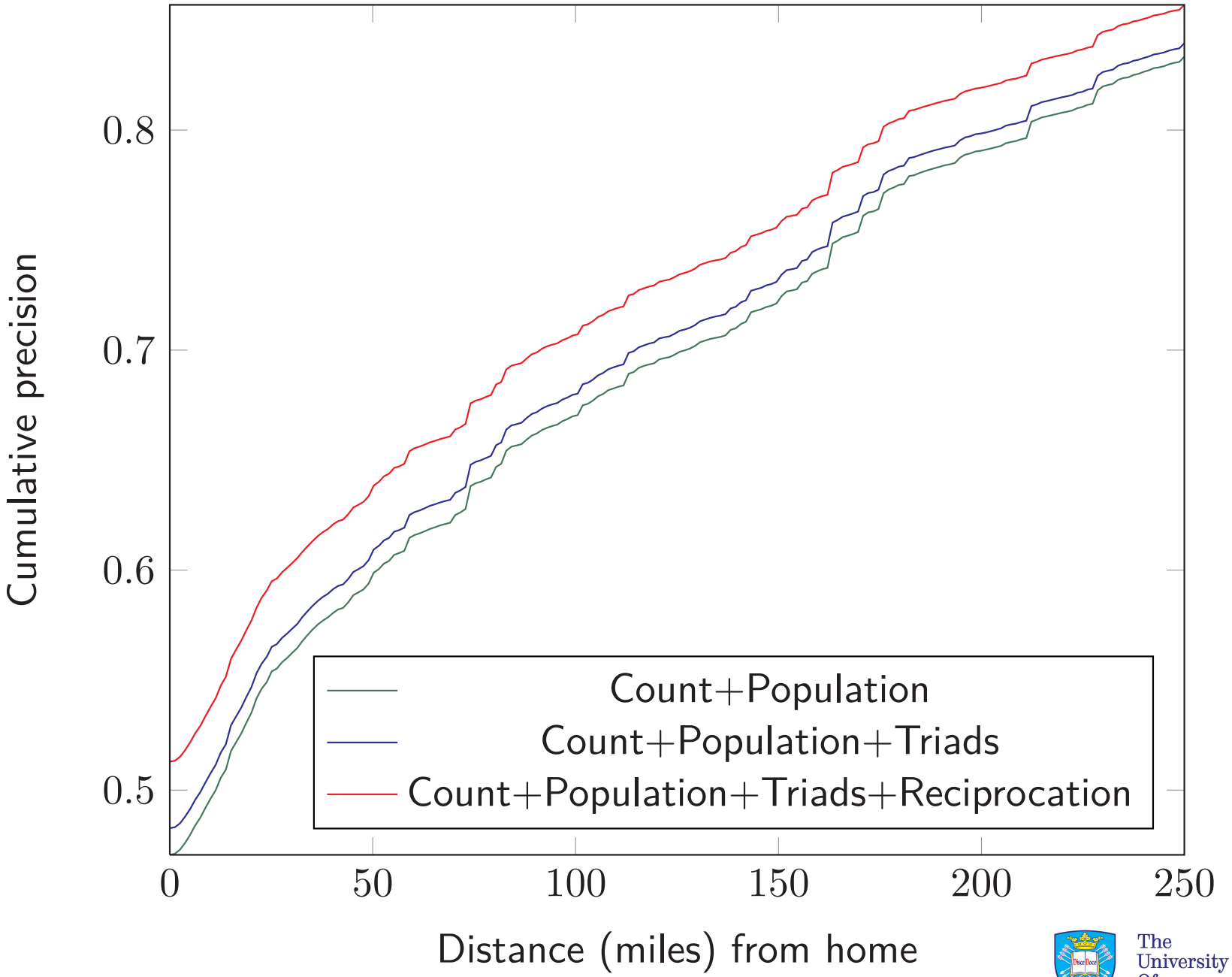


Forming Triads

People tend to befriend friends-of-friends over time.

1. Wally meets Jeff
2. Jeff meets (and possibly falls for) Sandra
3. Jeff introduces Sandra to Jeff





The University Of Sheffield.

Results

- All learned approaches outperformed baseline
- Locate user to within 100 miles at $69.03\% \pm 0.43\%$ accuracy

Method	≤ 0 mi	≤ 50 mi	≤ 100 mi
Random assignment	31.61%	43.61%	52.71%
Oracle performance	78.15%	89.59%	92.97%
Simple Friendship Count	39.49%	47.79%	55.18%
Inverse City Frequency	10.66%	40.59%	51.34%
Population	45.94%	58.20%	65.44%
Population & Triad	47.13%	59.24%	66.39%
Pop. & Triad & Recip.	**50.08%**	**62.08%**	**69.03%**

The University Of Sheffield.

Efficiency

- Our data collection for this task was very straightforward
  - Around 2 API requests per user to download graph
  - List of users taken from Spritzer Streaming API
- We gathered the social network in weeks
- Text of tweets took months
- Easier to apply in real-time than methods based on text of tweets

Data & Future Work

As part of this work, we developed a data set of users and associated geographical locations for the UK.

- This data is available anonymised online `http://www.domrout.co.uk`
- Data set has already been re-used in Lampos et al (2013) - modelling voter intentions.

In our future work, we hope to:

- Incorporate textual posts from users in our classification setting
- Explore more graphical features (such as betweenness of a node)
- Generalise decoding of user profile strings to locations worldwide

Thank you for listening!

The University Of Sheffield.