Inferring the location of Twitter messages based on user relationships

Clodoveu Davis
Gisele L. Pappa
Diogo Rennó R. de Oliveira
Filipe de L. Arcanjo

Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Geocoding Social Networks

- Social networks are becoming a valuable source of information
  - What is going on *right now*?
  - How fast is an idea/joke/trend/news spreading?
  - What is the source of a rumor?
  - What can quickly capture the attention of many people?

twitter blog

Location, Location, Location

Thursday, August 20, 2009

How To Tweet With Your Location

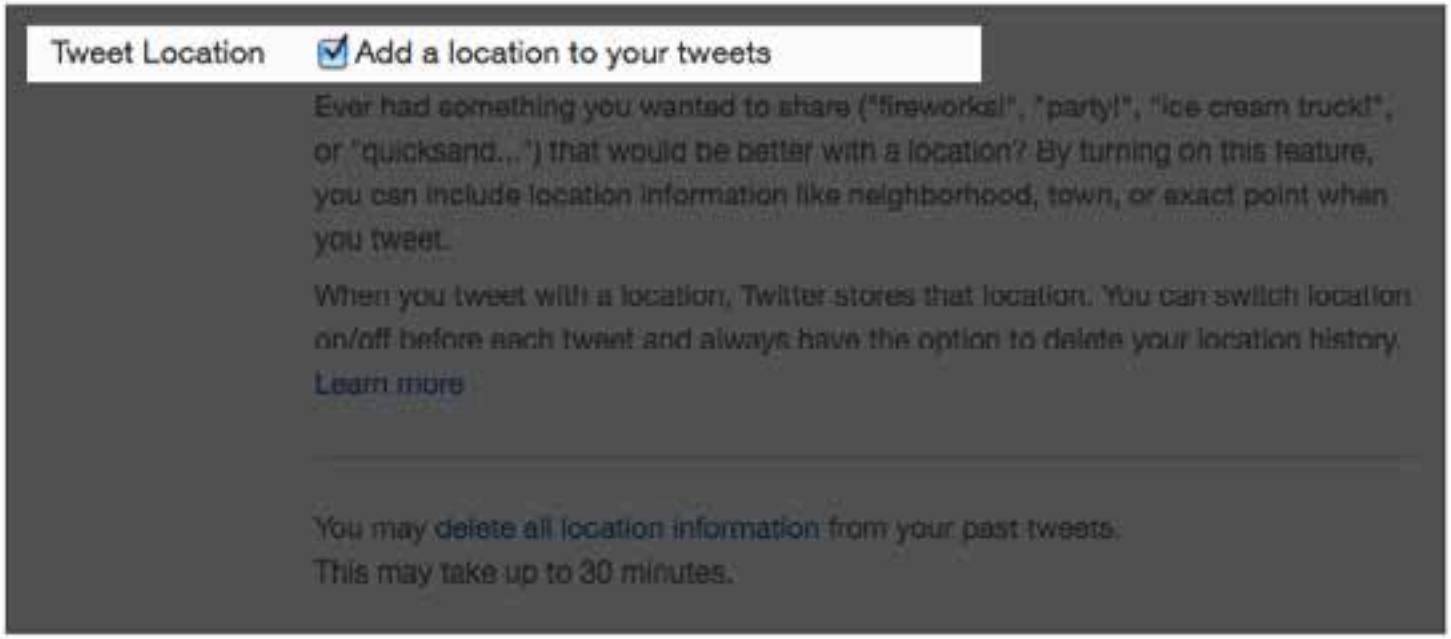
Tweet

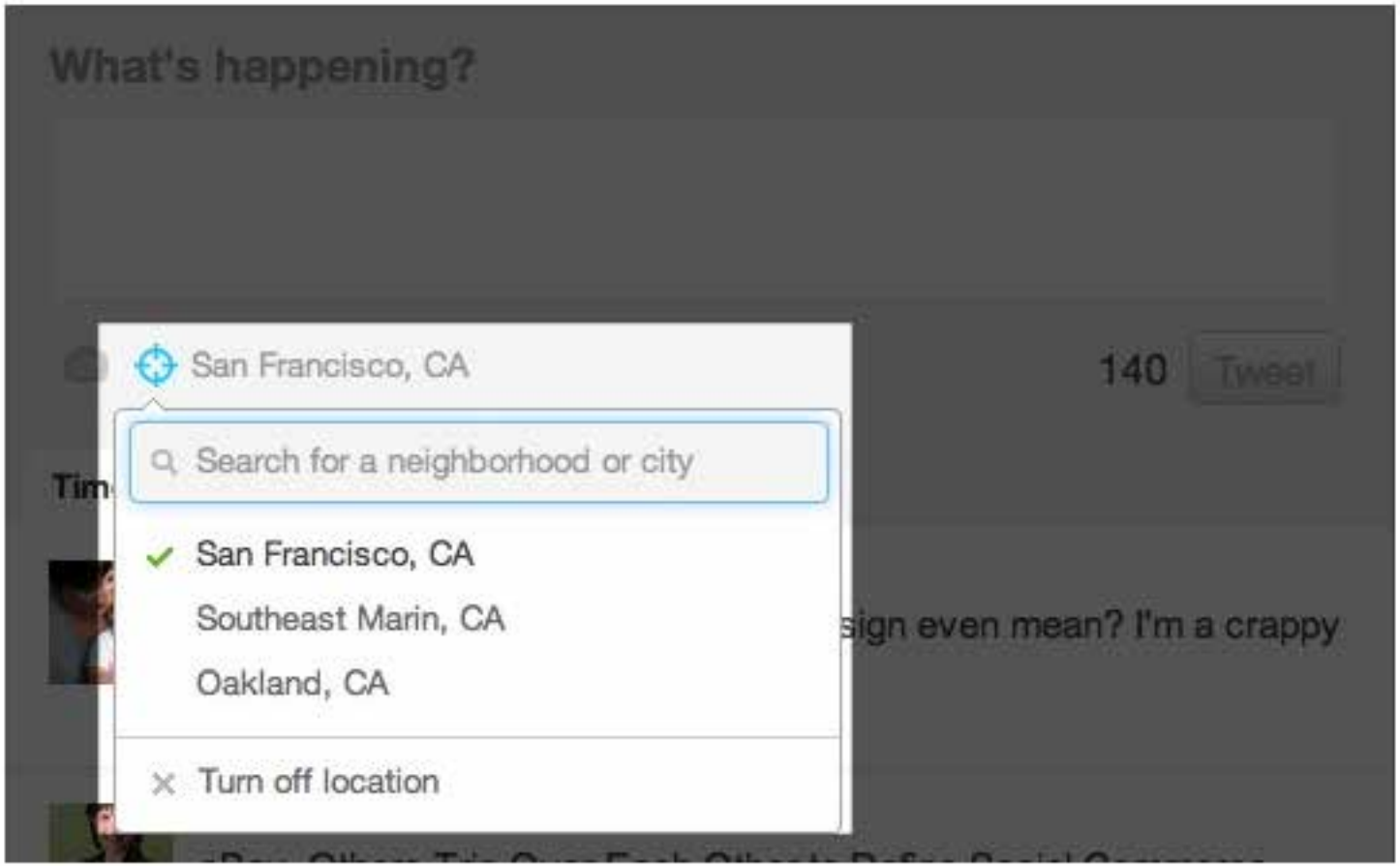
If you are not sure what tweeting with your location is, please visit our About the Tweet Location Feature help page for more information.

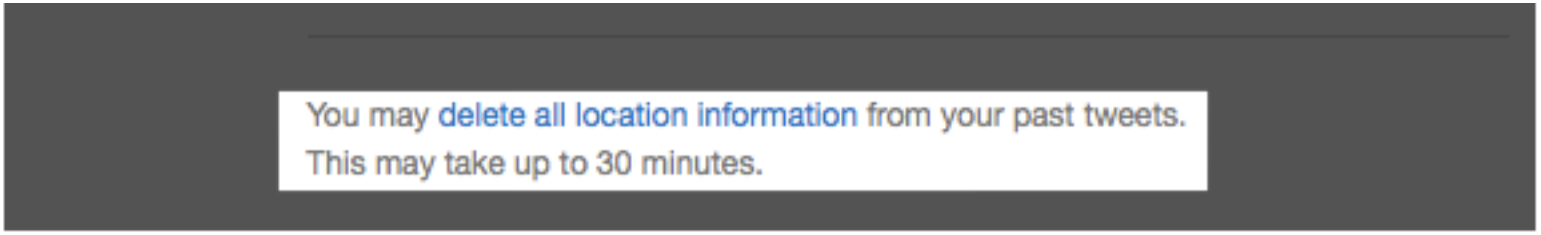
How to enable tweeting with your location:

Because tweeting with your location is disabled by default for everyone, you must enable this feature by following these steps:

1. Click on your Account Settings.
2. At the bottom of the page, check the box to "Add a location to your Tweets", as shown below.
3. Save your changes.




What's happening?
San Francisco, CA
140
Tweet
Search for a neighborhood or city
San Francisco, CA
Southeast Marin, CA
Oakland, CA
Turn off location
sign even mean? I'm a crappy


You may delete all location information from your past tweets.
This may take up to 30 minutes.

observatóriodaweb

Observatório da Web

Com aproximadamente 65 milhões de brasileiros conectados à Internet – o correspondente a 36% da população, segundo dados do Comitê Gestor da Internet no Brasil – eventos como a campanha presidencial de 2010 começaram a refletir um fenômeno já observado em países desenvolvidos. Cada vez mais, a Internet exerce um papel relevante na formação da opinião pública. Com a proposta de acompanhar esta nova realidade, o Observatório da Web é uma ferramenta gratuita dedicada ao monitoramento de importantes fatos, eventos e entidades na rede mundial de computadores em tempo real.

Observatórios

Três versões do Observatório da Web já foram construídas:
Observatório da Dengue | Observatório das Eleições 2010 | Copa do Mundo de Futebol 2010 |

observatóriodadengue

O Observatório da Dengue é um sistema de vigilância epidemiológica ativa a partir de dados internet, desenvolvido em parceria com o Instituto Nacional de Ciência e Tecnologia em dengue (INCT em dengue). O Observatório da Dengue é capaz de coletar, analisar e apresentar em tempo real informações acerca da dengue a partir de mais de uma centena de fontes de dados internet, incluindo redes sociais e blogs, além de canais da mídia tradicional. O sistema permite visualizar as informações coletadas de diversas formas e provê estimativas acerca da incidência da dengue em determinada região em tempo real, sem o atraso que ocorre quando há a necessidade da notificação e entrada de dados no sistema tradicional de controle epidemiológico. As estimativas se baseiam na correlação espaço-temporal da incidência de dengue entre dados obtidos da internet e os dados gerados pelo sistema de notificação do Ministério da Saúde.
Visitar o site

Tecnologia

Os pesquisadores da UFMG utilizam um conjunto inédito de tecnologias de engenharia na Web - como recuperação de informação, gerenciamento de dados da Web, mineração de dados e visualização – para entender o que está sendo veiculado nas várias mídias e pelos vários usuários. Esse entendimento é fundamental para avaliar o efetivo impacto das campanhas na Internet e como os usuários interagem e reagem às notícias e discussões.

A consulta à visibilidade no Twitter, rede social com crescente popularidade no país, traz outro recurso diferenciado além da nuvem de tags: a propagação dos tweets. "Somos os primeiros a mostrar quantas pessoas foram atingidas por uma mesma informação em um intervalo de tempo", destaca Wagner Meira, pesquisador do INWeb.

Projeto INWeb

O Observatório da Web integra o Observatório da Web, um dos projetos de pesquisa do Instituto Nacional de Ciência e Tecnologia para a Web, financiado pelo CNPq e pela Fapemig. Colaboram com o estudo cerca de 30 especialistas de quatro instituições federais de ensino: Universidade Federal de Minas Gerais (UFMG), Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Universidade Federal do Amazonas (UFAM) e Universidade Federal do Rio Grande do Sul (UFRGS).

Instituições
Departamento de Ciência da Computação - UFMG
InWeb
CNPq

Informações
Contatos

Comparativos

1- Escolha a(s) personalidade(s) para a análise e informe o período:


Veja também

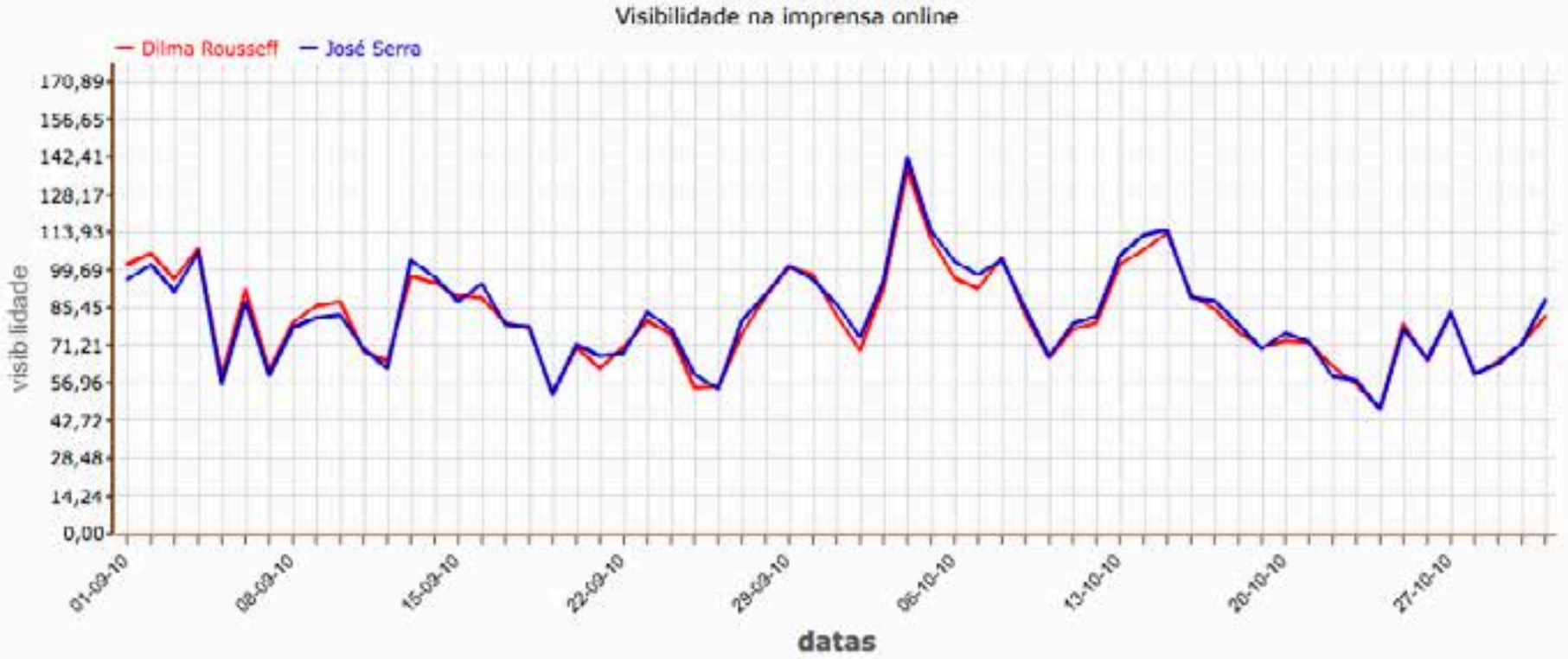
Comparativos

Visibilidade

Citações

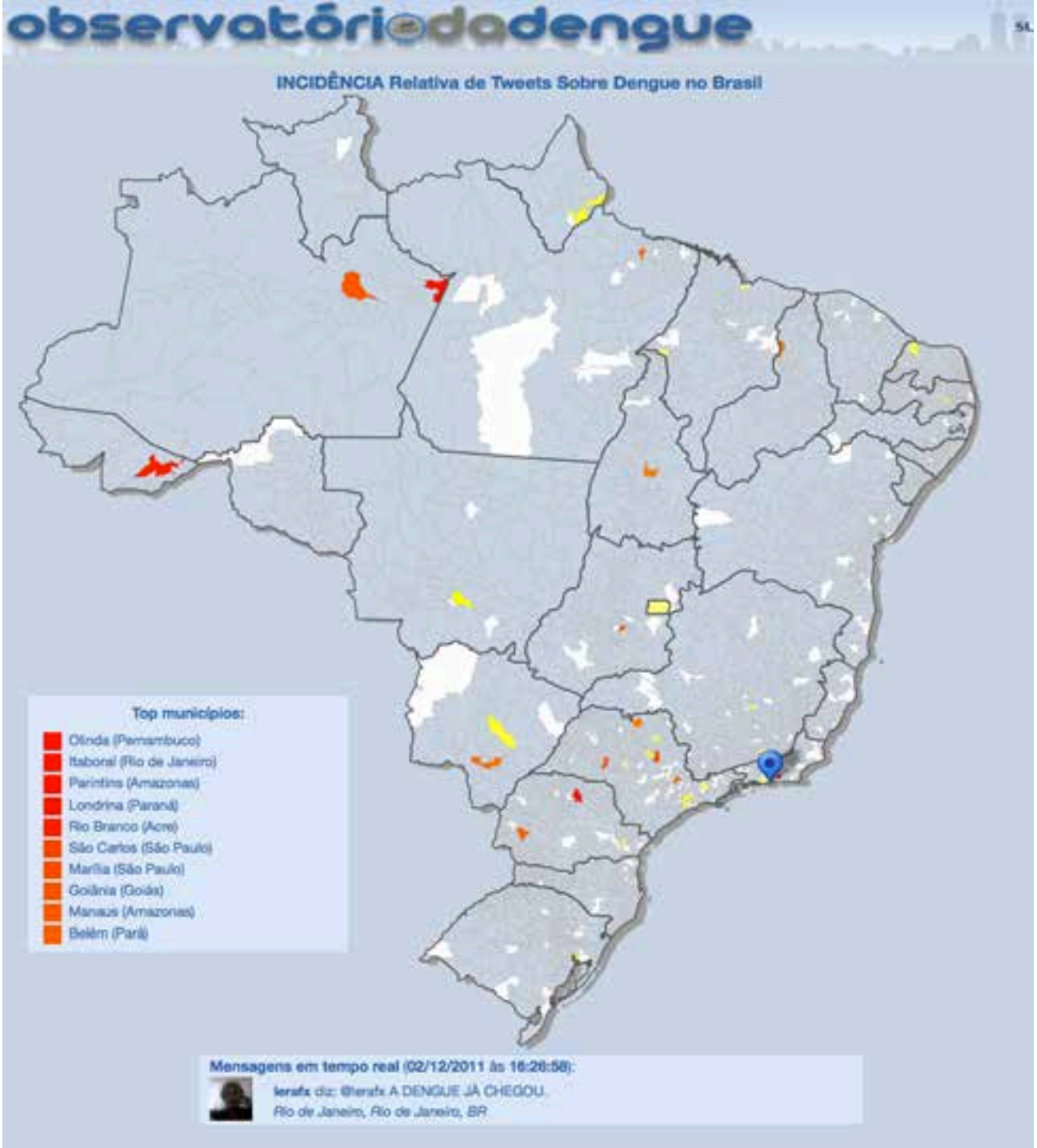
Selecione o período: Últimos dois meses (Nota: o período é referente à data 31-10-2010)

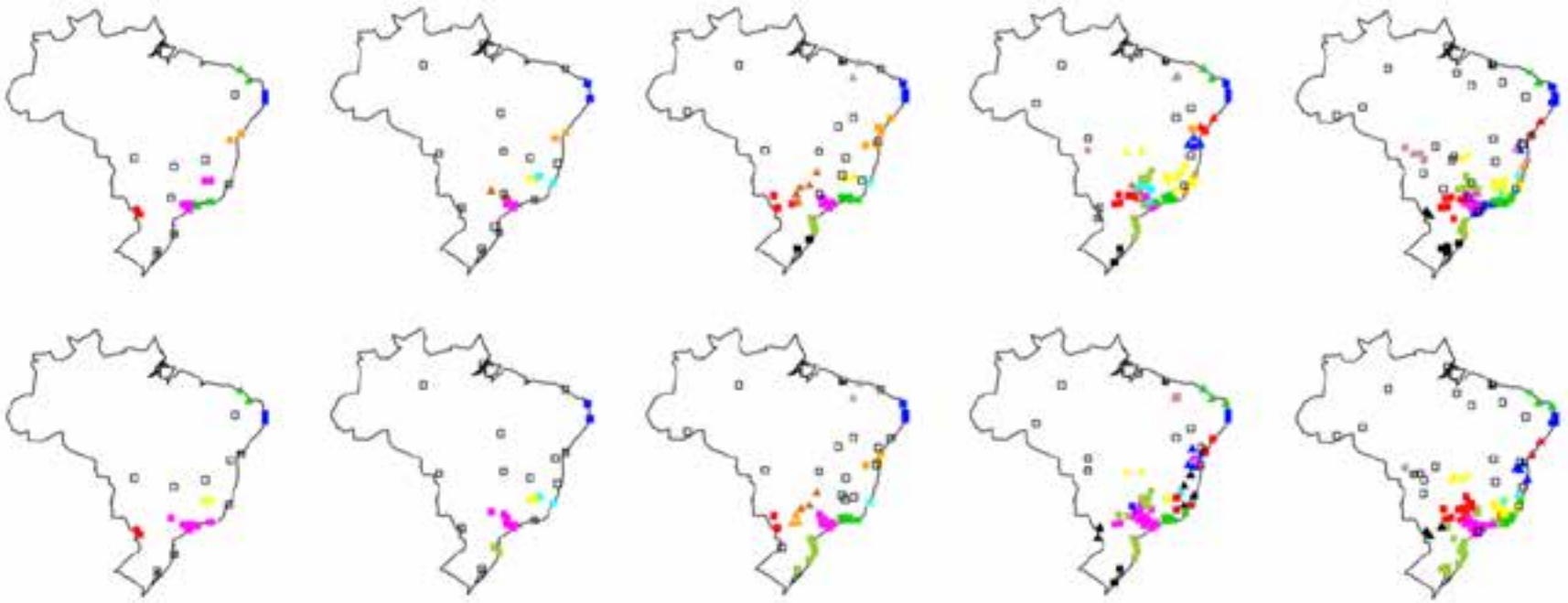
Ok, mostre-me os resultados



Como esta análise é construída

O índice de visibilidade não se refere ao número absoluto de citações dos candidatos. Trata-se de uma pontuação desenvolvida para evitar distorções levando em


observatóriodadengue
INCIDÊNCIA Relativa de Tweets Sobre Dengue no Brasil
Top municípios:
Olinda (Pernambuco)
Itaboral (Rio de Janeiro)
Parintins (Amazonas)
Londrina (Paraná)
Rio Branco (Acre)
São Carlos (São Paulo)
Marília (São Paulo)
Goiânia (Goiás)
Manaus (Amazonas)
Belém (Pará)
Mensagens em tempo real (02/12/2011 às 16:26:58):
lerafx diz: @lerafx A DENGUE JÁ CHEGOU.
Rio de Janeiro, Rio de Janeiro, BR



Clusters of cities with similar dengue rates

Top line: percentage of personal experience tweets on dengue
Bottom line: data from official dengue reports

Period: January to May 2009
Source: Gomide et al., 2010

Objectives

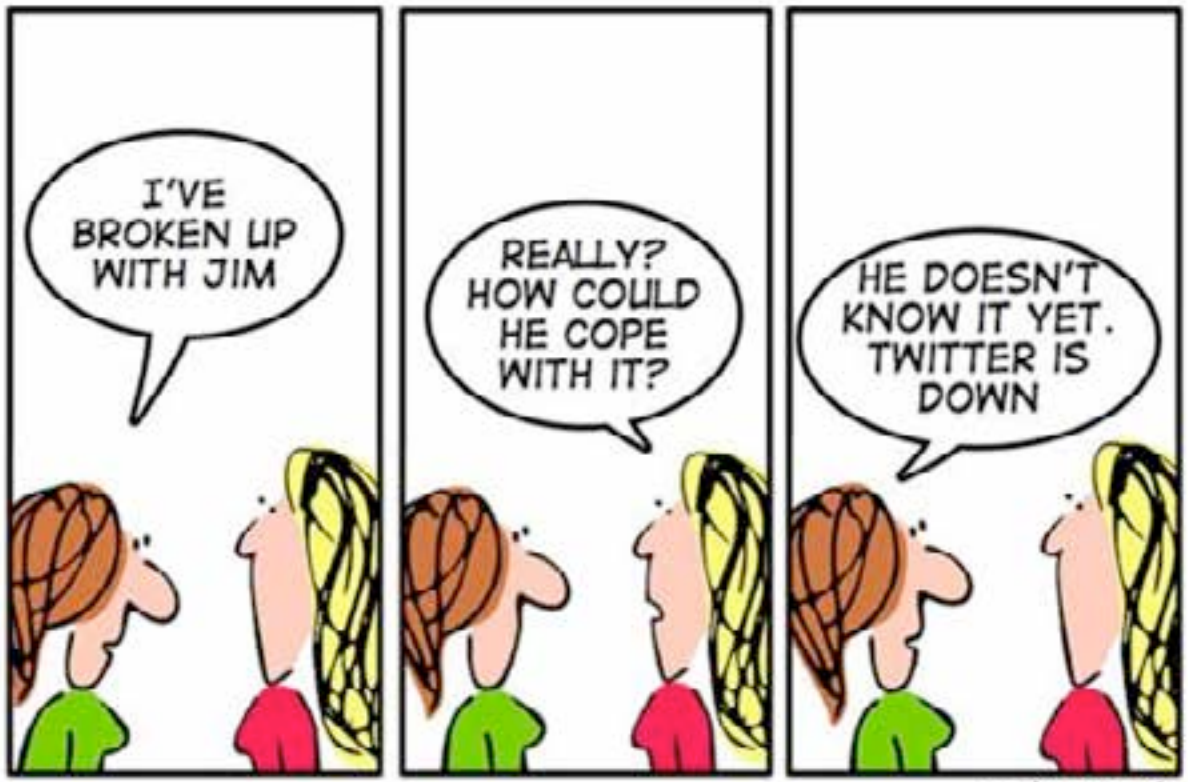
- Increase the number of locatable tweets
- Improve the possibility of tweet monitoring for relevant subjects in near real-time
  - Early warning system
- Adequately capture spatiotemporal trends in Twitter message streams

Idea

- Online friends usually live nearby
  - This has been demonstrated by previous works for Facebook
- Declared locations are unreliable and often fail to guide an accurate positioning
- Assume that unlocatable tweets by friends of someone whose location can be approximately established can be placed at that location as well

“Friends” in Twitter

- Twitter relationships are unilateral (*following* someone), not reciprocal as in Facebook or others

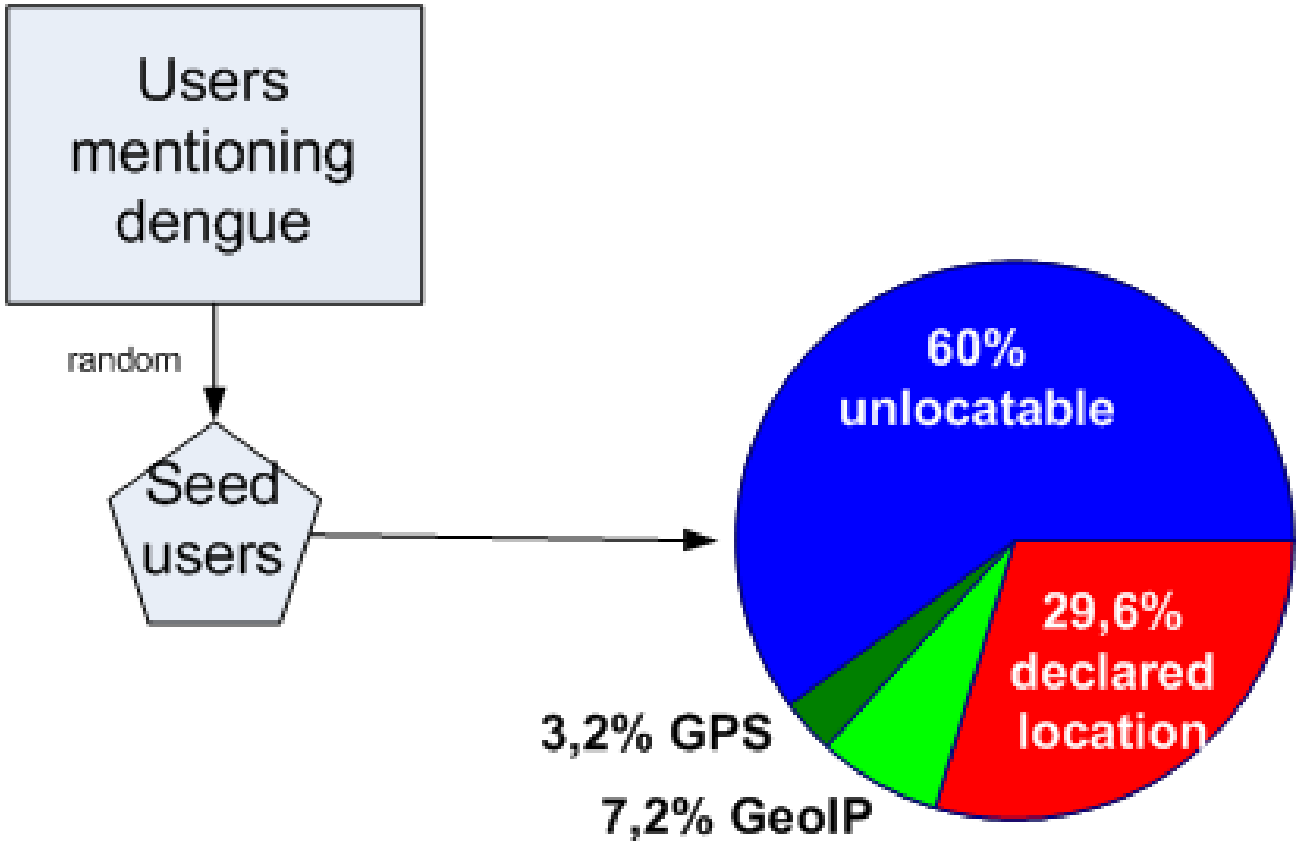


http://buncee.com/files/uploads/image/twitter-comic-2.jpg

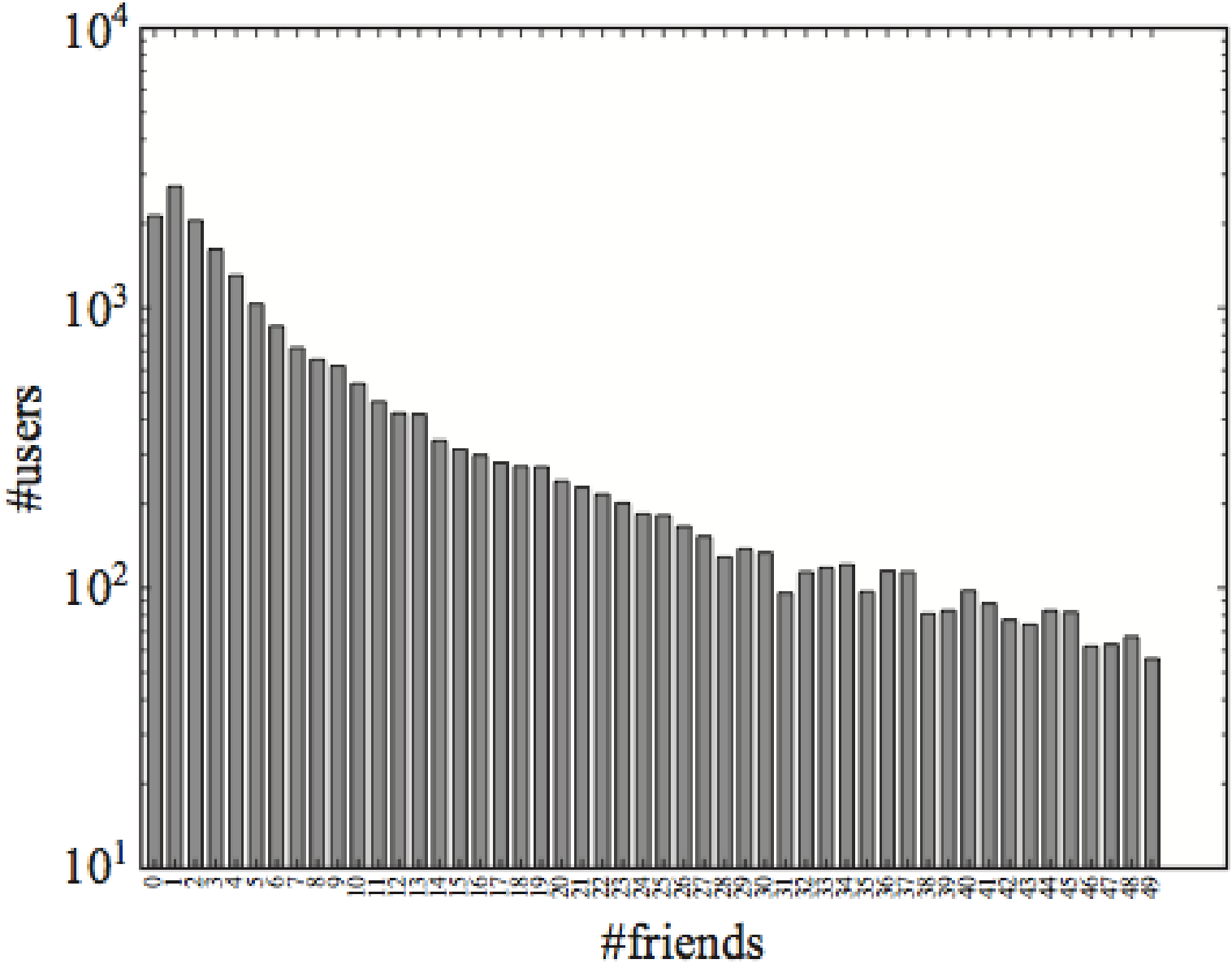
“Friends” in Twitter

- Only 22% of Twitter relationships are reciprocal
  - We consider those to be indicative of friendship
  - Intersection between the sets of *followers* and *followed_by* of each user

Data collection



Valid location: a city name in Brazil



Users with a valid location and friends with valid location

Average precision and recall over a 10-fold cross validation

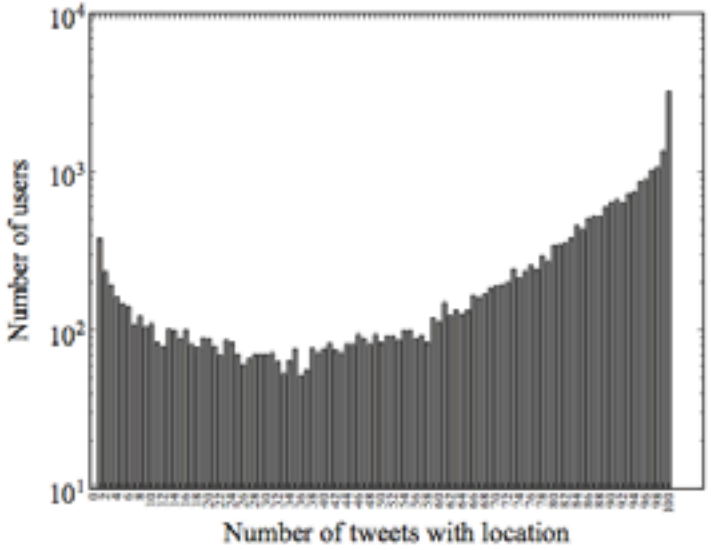
Data	No. of users	Precision	Recall
All	24,767	0.4013 ± 0.0042	0.9137 ± 0.0046
GPS	566	0.3262 ± 0.0776	0.3440 ± 0.0341
GeoIP	1,606	0.2178 ± 0.0289	0.4387 ± 0.0266
Declared	22,595	0.4047 ± 0.0050	0.9103 ± 0.0043

Simple voting scheme
No recall = no friends or tied voting

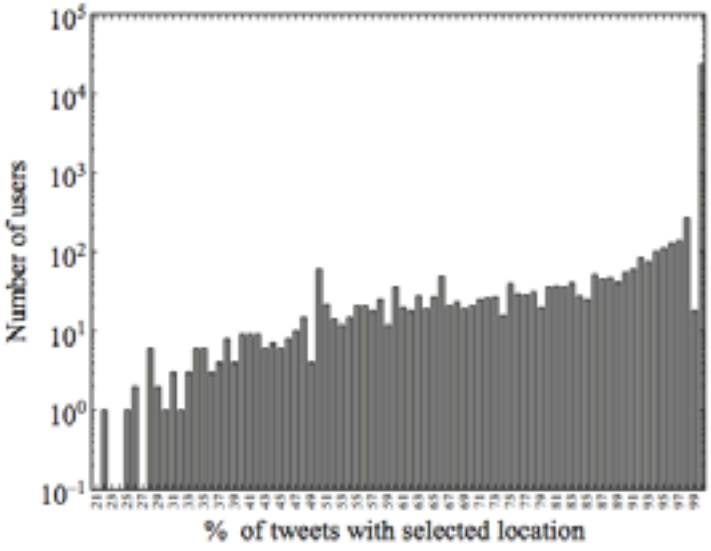
Location Inference

- Use the most frequent location amongst friends
  - If the user has too few friends at the same location, this can lead to distortions
- Parameters
  - Minimum and a maximum number of locatable friends
  - Minimum number of votes required for a location to be considered trustworthy
- Evaluation using 10-fold cross validation

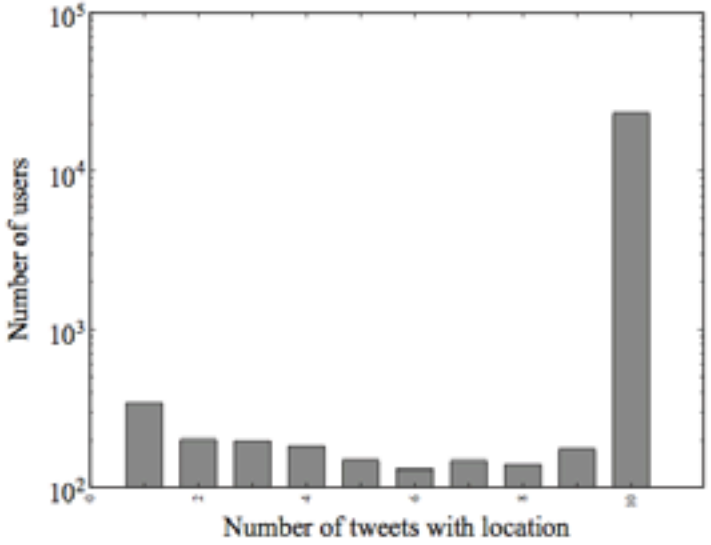
Tweets with location



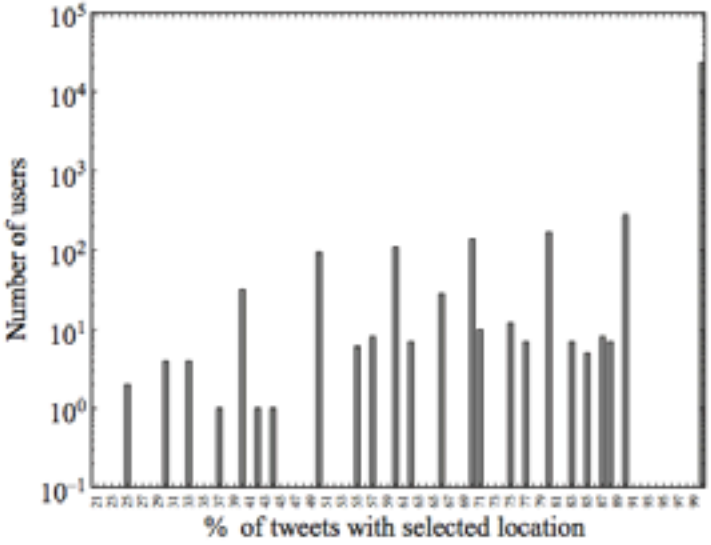
(a) Number of tweets with location (max. 100)



(b) % of tweets with MFL

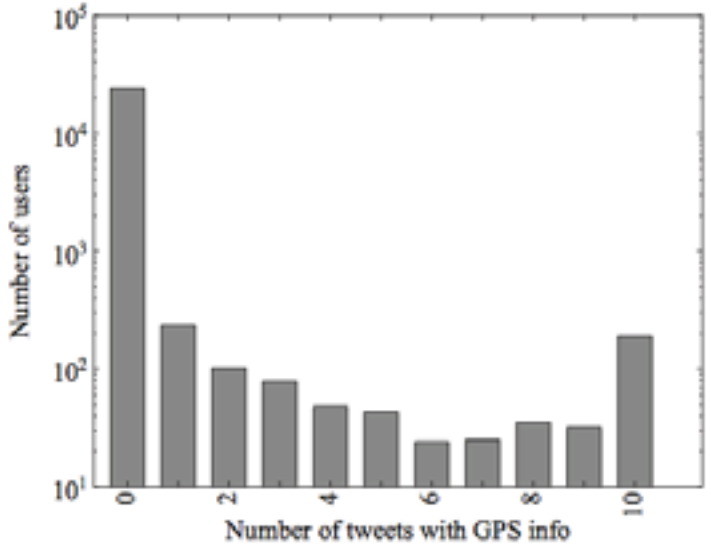


(c) Number of tweets with location (max. 10)

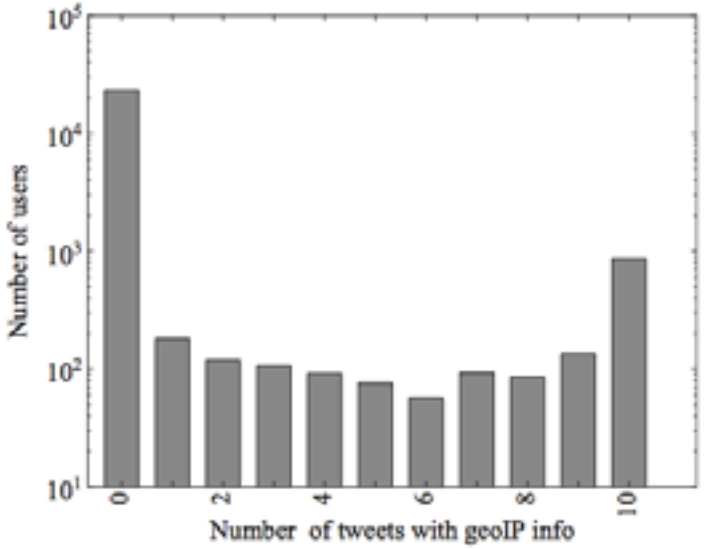


(d) % of tweets with MFL

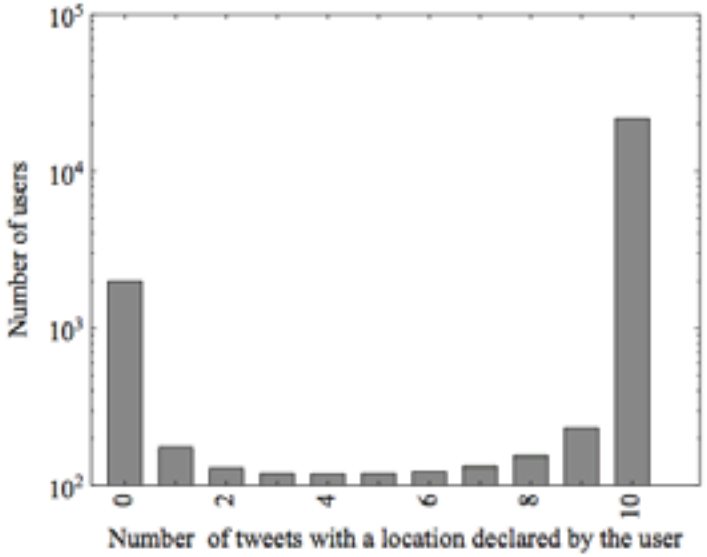
Source of location data



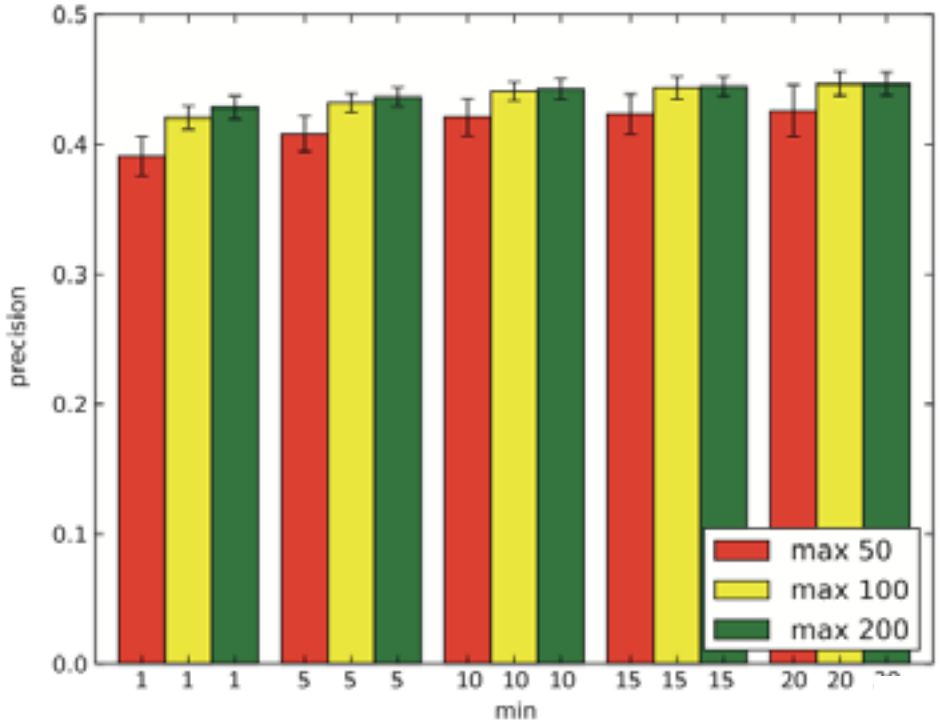
(a) GPS Location



(b) geoIP Location



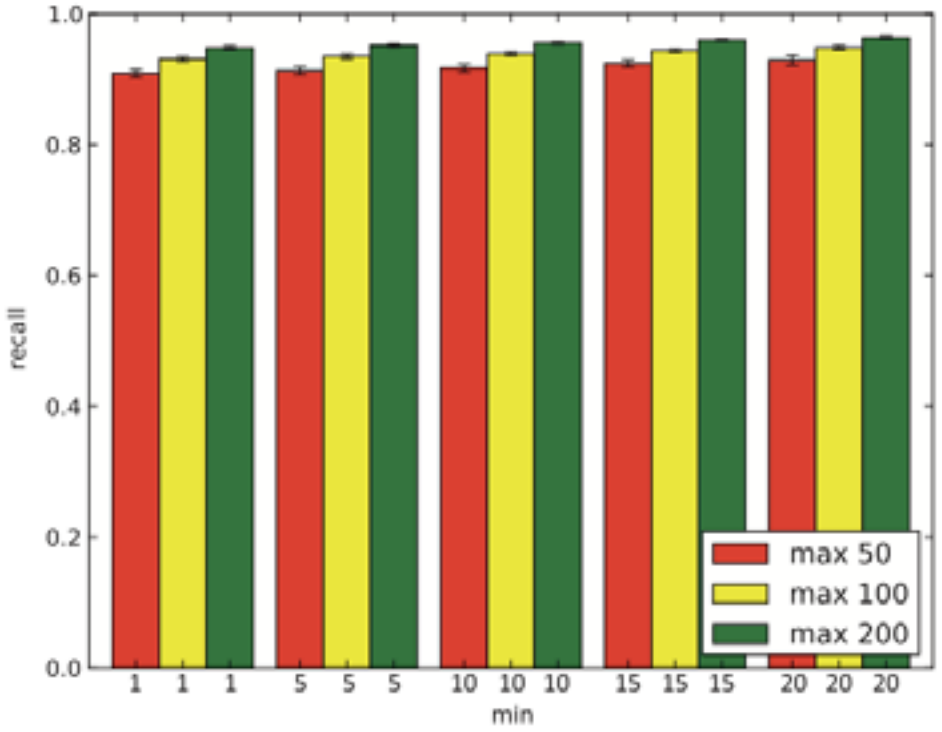
(c) User-provided location

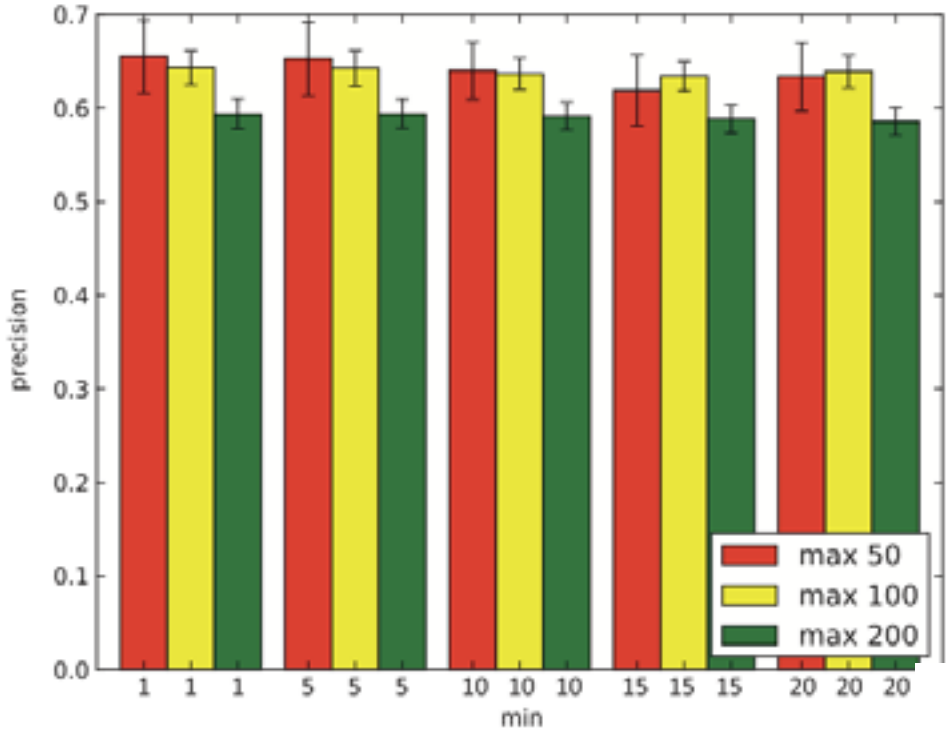


Precision and recall: users with # of friends between min and max

Intervals have little influence over precision (max 44%)

Recall increases as the maximum increases (~90-95%)

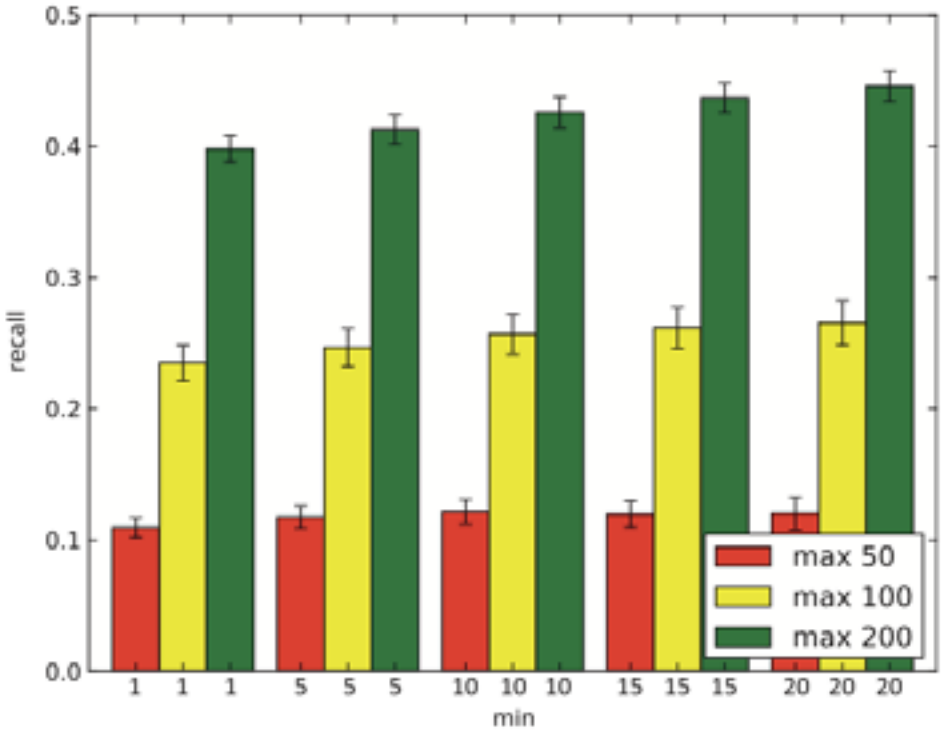


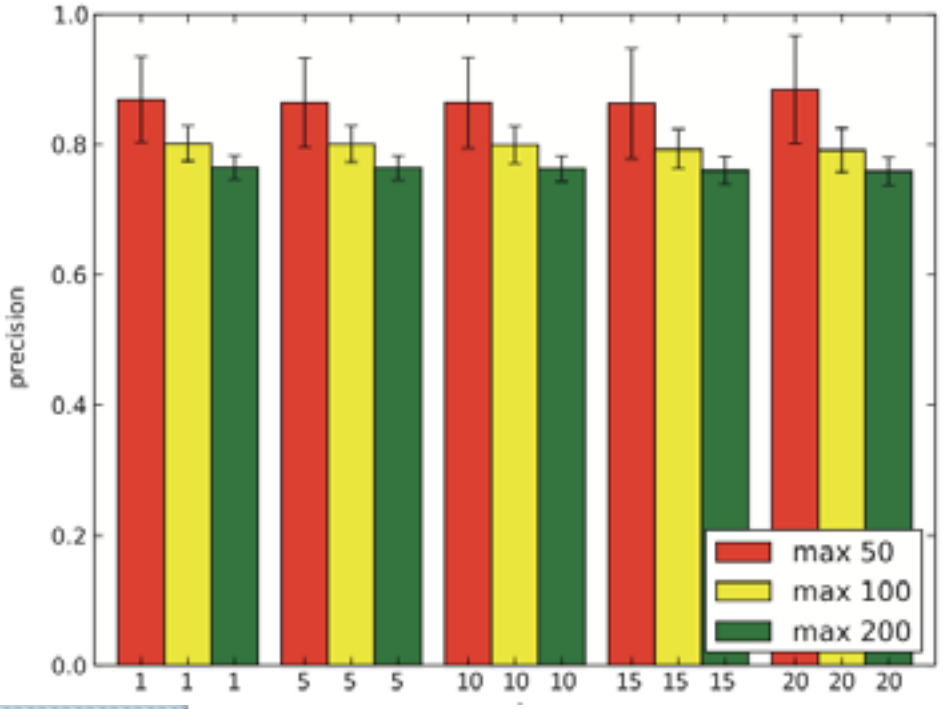


Precision and recall: users and their friends with # of friends between min and max

Intervals again have little influence over precision, but it is higher than before (65%)

Recall increases as the maximum limit increases (~40-45%)

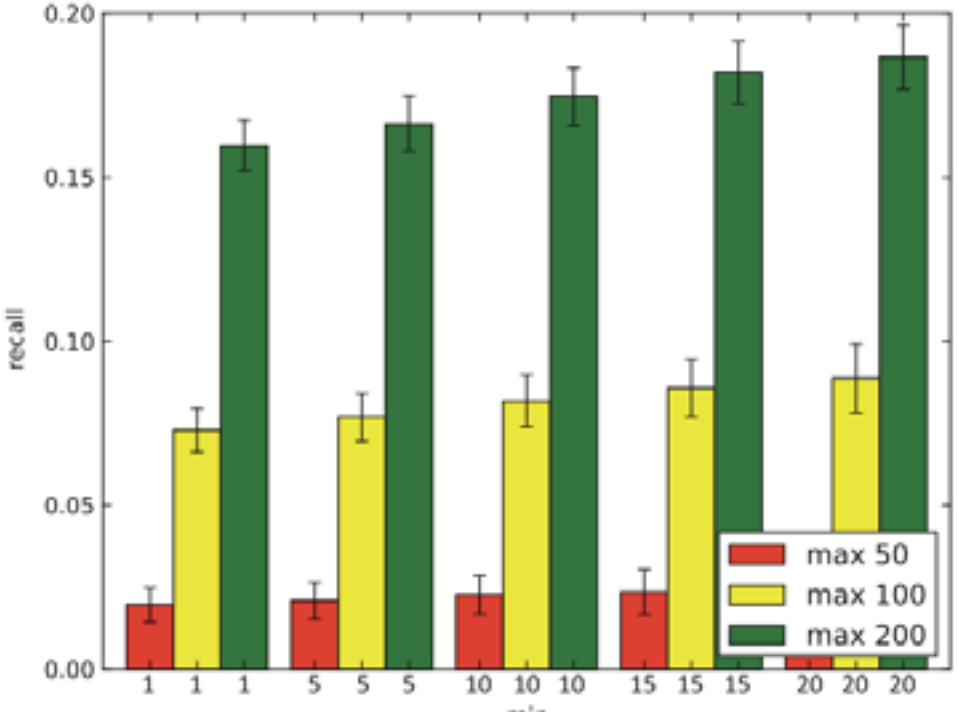




Precision and recall: min. 2 votes by users who have a # of friends between min and max

As expected, the greater the interval, the lower the precision and the higher the recall

A recall of 18% with 80% accuracy represents a 45% improvement over the original 40% locatable tweets



Conclusions

- Using large numbers of personal messages allows assessing the repercussion of events and detecting trends
- Knowing the geographic origin and reach of this repercussion is very important
- Inference method can improve the number of locatable tweets by about 45%
- Too few or too many relationships can increase uncertainty

Future work

- Implement more sophisticated method for location inference
  - Bayesian networks
  - Spreading activation
  - Fuzzy propagation in graphs
- Extend to other social networks
- Increase spatial granularity (intra-urban)
- Apply to other phenomena and subjects of popular interest

clodoveu@dcc.ufmg.br