

Técnicas de Coleta

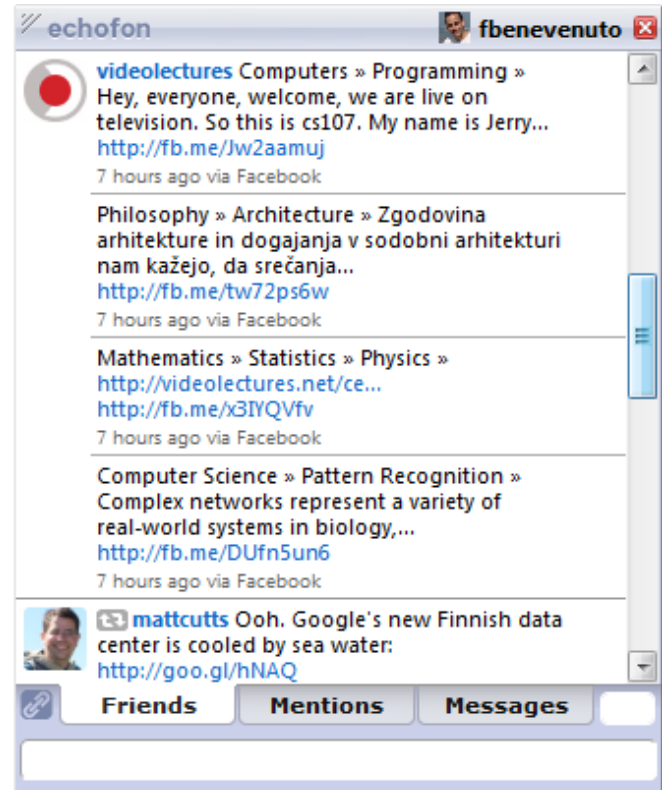
Ana Paula

Fabrizio Benevenuto

API do Twitter

- Permitem a construção de aplicações, mas podem ser utilizadas por crawlers

- **statuses/filter**
- **statuses/sample**
- **trends**
- **trends/daily**
- **trends/weekly**
- **statuses/retweets_of_me**
- **statuses/mentions**
- **account/rate_limit_status**



API do Twitter

- Profile do usuário: <http://twitter.com/users/show/44446416.xml>

```
- <user>
  <id>44446416</id>
  <name>Fabricio Benevenuto</name>
  <screen_name>fbenevenuto</screen_name>
  <location>Belo Horizonte - Brazil</location>
- <description>
  PhD candidate at Federal University of Minas Gerais.
</description>
- <profile_image_url>
  http://a3.twimg.com/profile_images/298811199/me_normal.jpg
</profile_image_url>
<url>http://www.dcc.ufmg.br/~fabricio</url>
<protected>>false</protected>
<followers_count>201</followers_count>
```

API do Twitter

- Tweets: http://twitter.com/statuses/user_timeline.xml?user_id=44446416&count=200&page=1

```
-<status>
  <created_at>Fri Jul 16 17:59:32 +0000 2010</created_at>
  <id>18704982149</id>
  -<text>
    No aeroporto preparando pra maratona de voos ate casa... Todos os voos na cadeira do meio e dessa vez tem até troca de aeroporto no Rio...
  </text>
  -<source>
    <a href="http://www.echofon.com/" rel="nofollow">Echofon</a>
  </source>
  <truncated>>false</truncated>
  <in_reply_to_status_id/>
  <in_reply_to_user_id/>
  <favorited>>false</favorited>
  <in_reply_to_screen_name/>
+<user></user>
  <geo/>
  <coordinates/>
  <place/>
  <contributors/>
</status>
```

API do Twitter

- Followees: Provê 5000 IDs por requisição
- <http://twitter.com/friends/ids/44446416.xml?page=1>

```
- <ids>  
  <id>52806725</id>  
  <id>683113</id>  
  <id>155308339</id>  
  <id>21339294</id>  
  <id>47725447</id>  
  <id>53961984</id>  
  <id>39665161</id>  
  <id>22594570</id>  
  <id>128580638</id>  
  <id>61744603</id>  
  <id>80429908</id>  
  <id>66700199</id>  
  <id>44885947</id>  
  <id>14252137</id>  
  <id>633</id>  
  <id>56399566</id>  
  <id>39615488</id>  
  <id>50999197</id>  
  <id>82782832</id>
```

API do Twitter

- Followers: <http://twitter.com/followers/ids/44446416.xml?page=1>

```
-<ids>  
  <id>169214931</id>  
  <id>52806725</id>  
  <id>130842043</id>  
  <id>54559992</id>  
  <id>22851900</id>  
  <id>108289344</id>  
  <id>17683185</id>  
  <id>144301571</id>  
  <id>162897056</id>  
  <id>162235061</id>  
  <id>89322379</id>  
  <id>20028008</id>  
  <id>155308339</id>  
  <id>29901018</id>  
  <id>53749745</id>  
  <id>68388685</id>  
  <id>153812691</id>  
  <id>17417486</id>  
  <id>14665249</id>
```

API do Twitter

- http://twitter.com/help/request_whitelisting

Request whitelisting

Please submit this form as the user you would like an increased/lifted rate limit for. Before you apply, review our [documentation on API rate limits](#). Whitelisting is **only** available to developers and to applications in production; **all other requests are rejected**.

Follower limits are **not** affected by API whitelisting. If you are hitting the follower limit, please consult [our support documentation](#). API whitelisting **will not solve your problem** in this case.

Finally, if any of this is confusing to you, then whitelisting is probably not the answer to your question or problem. Please [visit our support site](#) to resolve your issue.

Do you want to whitelist an IP(s) in addition to your account?

List the address or addresses below as [CSV](#). IP ranges and netblocks are not accepted.

Describe your project in detail

Specify the methods you'll be using, the functionality of your application, and the expected frequency of use.

Please provide contact information

Specify email addresses or phone numbers where we can contact you in case of emergency.

Introducing Twitter Data Grants

Wednesday, February 5, 2014 | By Raffi Krikorian (@raffi) [23:29 UTC]

Tweet

Today we're introducing a pilot project we're calling [Twitter Data Grants](#), through which we'll give a handful of research institutions access to our public and historical data.

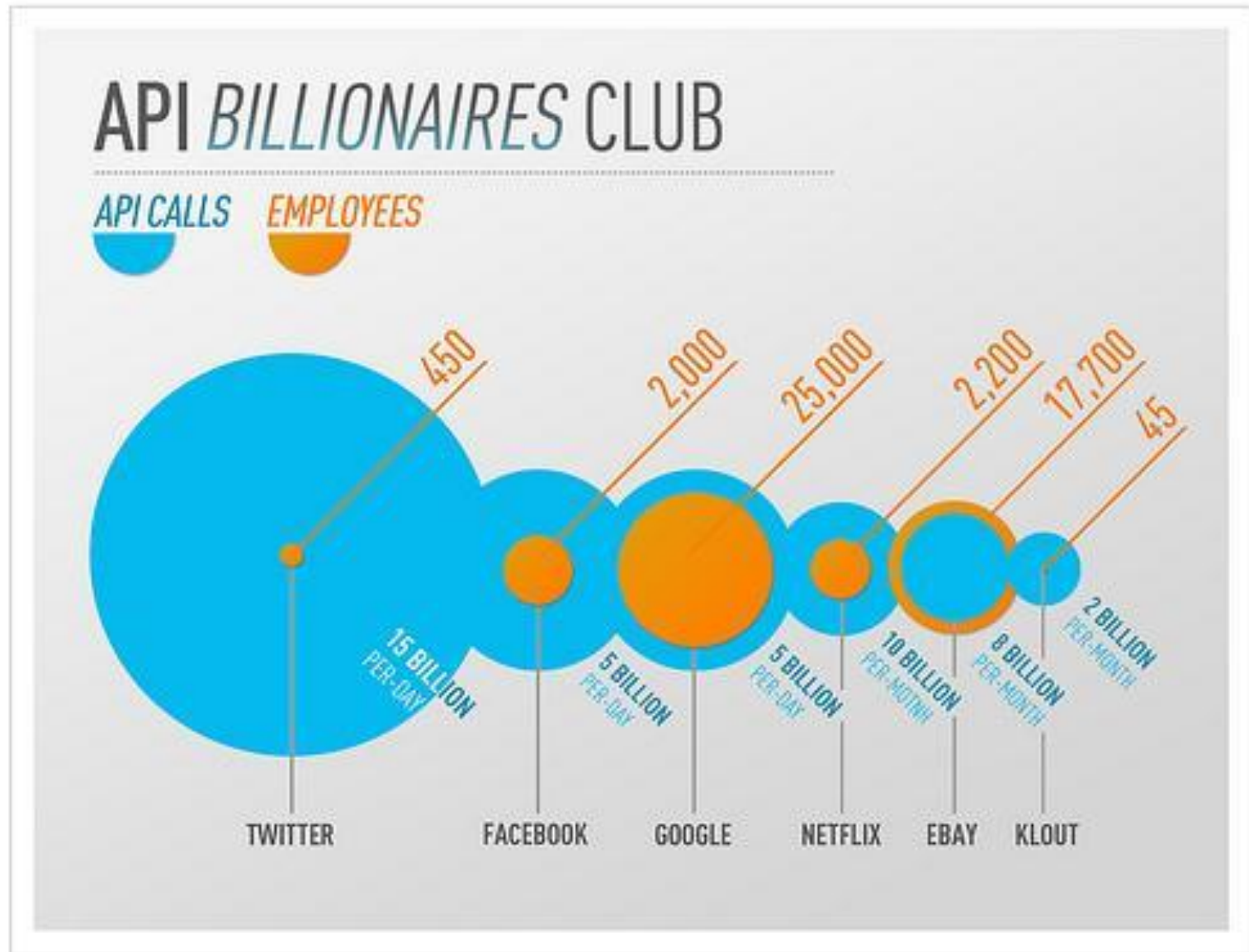
With more than 500 million Tweets a day, Twitter has an expansive set of data from which we can glean insights and learn about a variety of topics, from health-related information such as when and [where the flu may hit](#) to global events like [ringing in the new year](#). To date, it has been challenging for researchers outside the company who are tackling big questions to collaborate with us to access our public, historical data. Our Data Grants program aims to change that by connecting research institutions and academics with the data they need.



Tweet

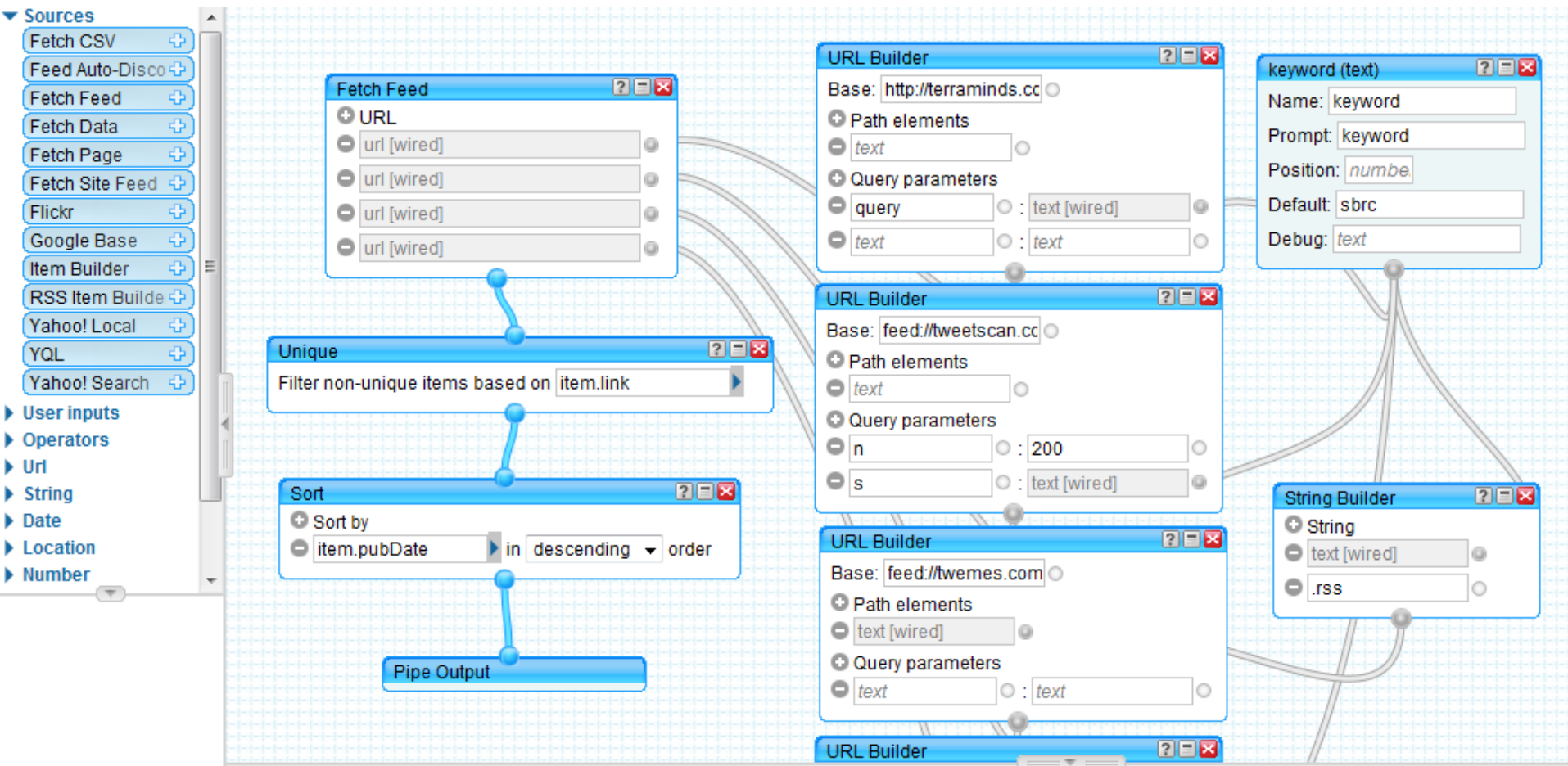
Submit a proposal for consideration to our [Twitter Data Grants pilot program](#) by March 15.

Popularidade das APIs



Mashups

- Aplicação que mistura várias APIs
- Yahoo Pipes: Pode ser útil também para coletar dados



Crawler – código em perl

- Biblioteca LWP da linguagem PERL

```
#!/usr/bin/perl
```

```
use LWP;
```

```
$ua = LWP::UserAgent->new();
```

```
$req = new HTTP::Request(GET => "http://twitter.com/friends/ids/44446416.xml?page=1");
```

```
$content = $ua->request($req)->content;
```

```
print "$content";
```

Crawler – código em perl

- Com mais detalhes no cabeçalho

```
#!/usr/bin/perl
```

```
use LWP;
```

```
$ua = LWP::UserAgent->new(cookie_jar => {}); #cookies
```

```
$ua->requests_redirectable(@list); # redirect
```

```
$useragentinfo = "Mozilla/5.0 (X11; U; Linux i686; de-AT; rv:1.7.2 Gecko/20040820 Debian/1.7.2-4)";
```

```
$ua->agent($useragentinfo . $ua->agent);
```

```
$req = new HTTP::Request(GET => "http://twitter.com/friends/ids/44446416.xml?page=1");
```

```
$content = $ua->request($req)->content;
```

```
print "$content";
```

Crawler – código em python

- Biblioteca urllib da linguagem PYTHON

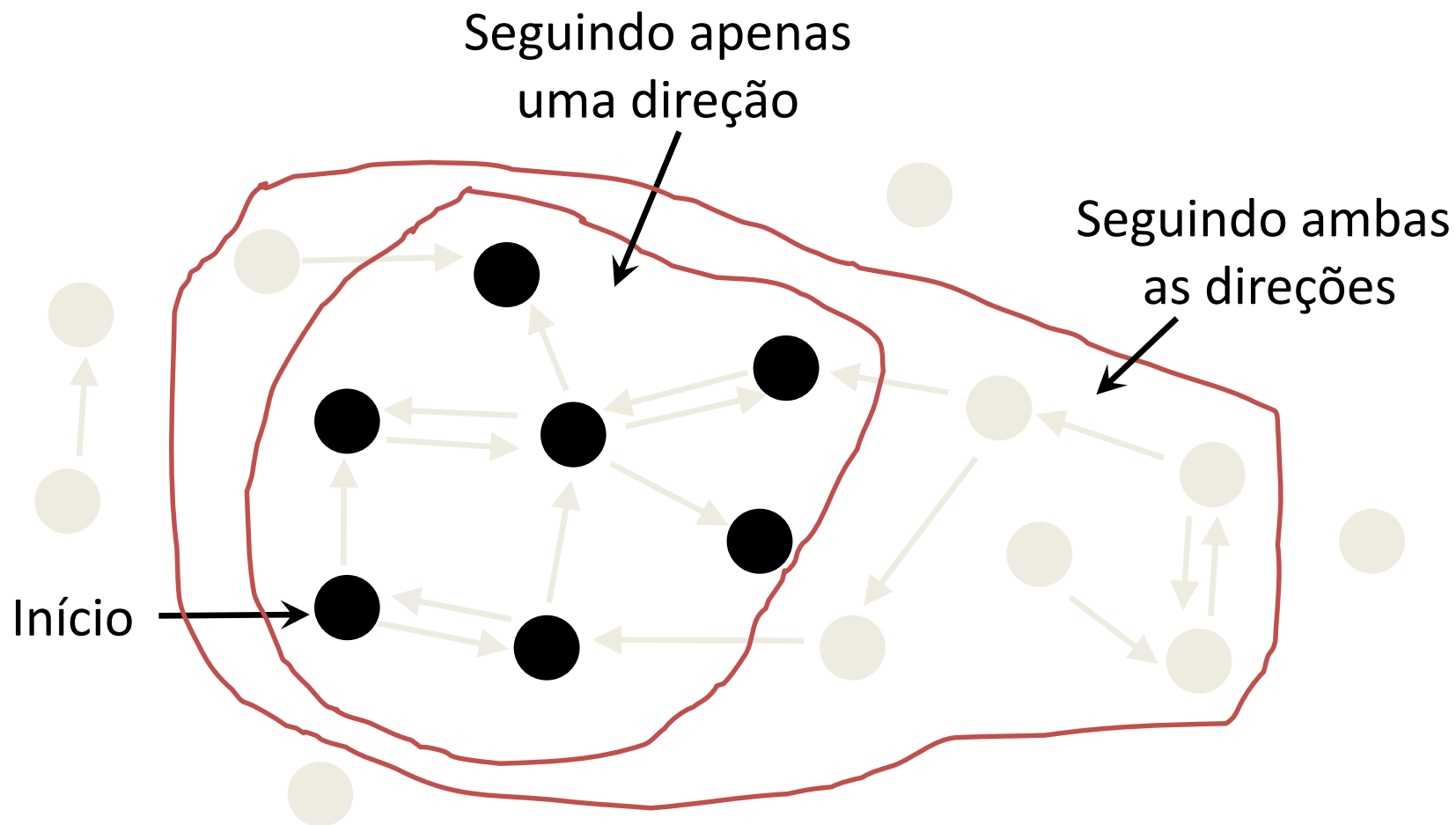
```
#!/usr/bin/python
```

```
import urllib
```

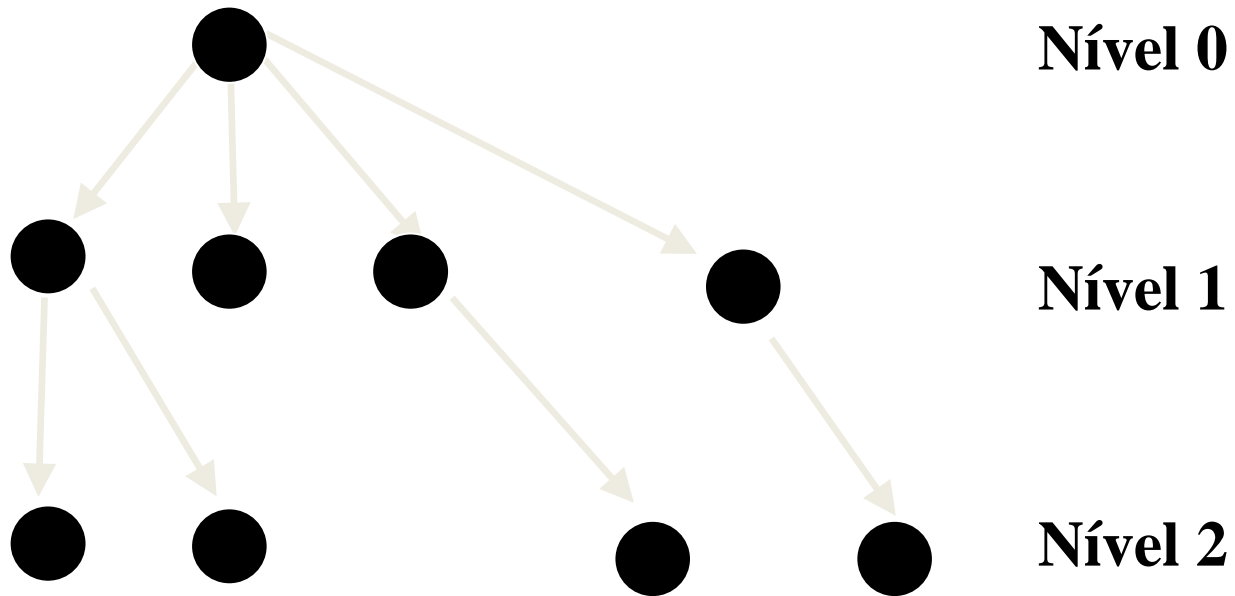
```
req = urllib.urlopen("http://twitter.com/friends/ids/44446416.xml?page=1")  
content = req.read()
```

```
print content
```

Coleta do WCC



Amostragem com Snowball



On the bias of BFS (Breadth First Search)

Maciej Kurant

School of Computer & Comm. Sciences
EPFL, Lausanne, Switzerland
maciej.kurant@gmail.com

Athina Markopoulou

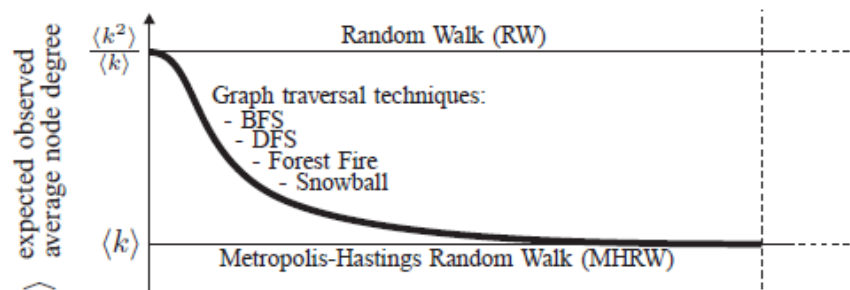
EECS Dept
University of California, Irvine
athina@uci.edu

Patrick Thiran

School of Computer & Comm. Sciences
EPFL, Lausanne, Switzerland
patrick.thiran@epfl.ch

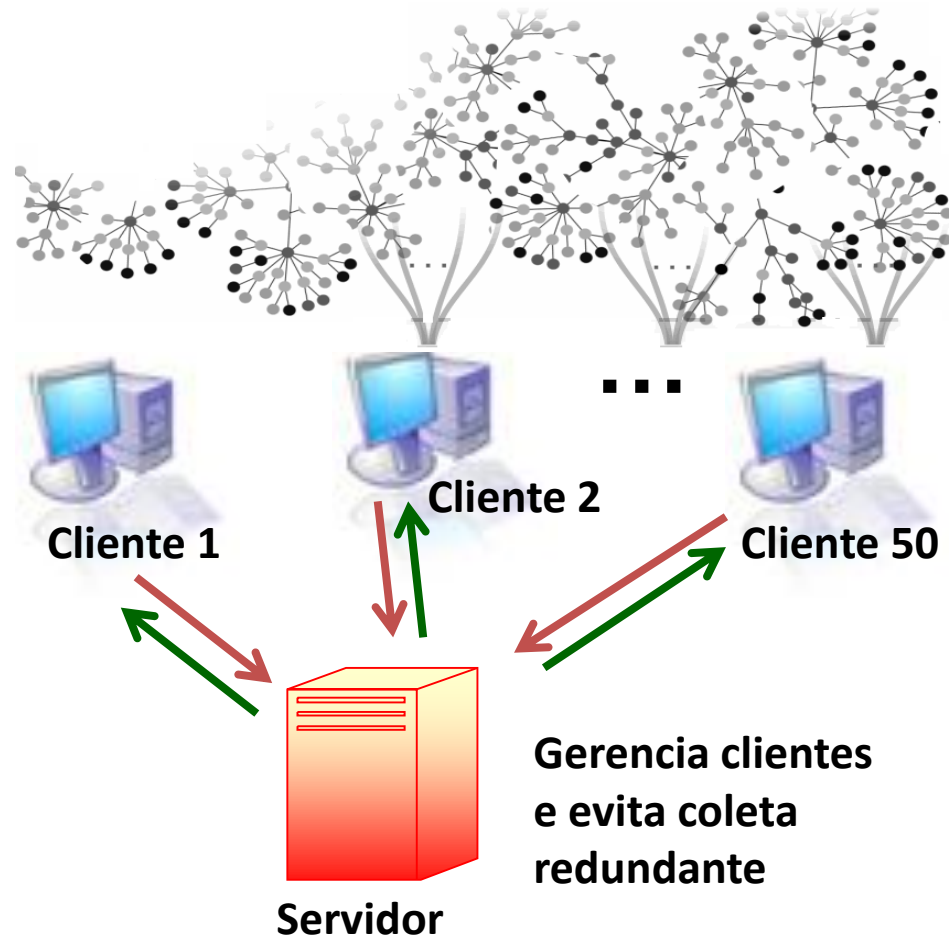
Abstract—Breadth First Search (BFS) and other graph traversal techniques are widely used for measuring large unknown graphs, such as online social networks. It has been empirically observed that incomplete BFS is biased toward high degree nodes. In contrast to more studied sampling techniques, such as random walks, the bias of BFS has not been characterized to date.

In this paper, we quantify the degree bias of BFS sampling. In particular, we calculate the node degree distribution expected to



Crawlers distribuídos

- **Clientes**
 - Recebem páginas do servidor para coletar
 - Coletam páginas
 - Encontram novas páginas a serem coletadas e devolvem ao servidor
- **Servidor**
 - coordena clientes
 - evita redundância
 - O servidor pode ser um simples banco de dados





Firebug

- JavaScript e Ajax muitas vezes escondem o HTML que procuramos com os crawlers.
- O Firebug é um add on do firefox que pode ajudar
- Qualquer ferramenta tcpdump like também pode resolver



Coletando o Orkut





ConsoleHTMLCSSScriptDOMNet

xhr

ClearPersistAllHTMLCSSJSXHRImagesFlash

URL	Status	Domain
+ http://www.orkut.com.br/Main#FriendsList?uid=8605703562113146391		
+ GET FriendsList?uid=86057035	200 OK	orkut.com.br



Console


HTML

CSS

Script

DOM

Net ▾



Clear

Persist

All

HTML



CSS

JS

XHR

Images

Flash

URL	Status	Domain
 GET Main#FriendsList?uid=860	200 OK	orkut.com.br
 http://www.orkut.com.br/FriendsList?uid=8605703562113146391		com.br



- IDs dos usuários são sequenciais no Twitter
 - Inspecionamos 80M de usuários, coletando perfil, todos os elos e tweets
 - Nenhum ID nas listas de seguidores/seguídos era superior a 80M
- Total de **55M de usuários, 2B de elos e 1.8B de tweets**
 - Cerca de 2 TB coletados
 - Lista branca para 58 máquinas no MPI-SWS
 - 20.000 requisições/hora em cada máquina
- Grafo de 55 milhões de nodos e 2 bilhões de arestas

ACM SIGCOMM WOSN 2009

Hot Today, Gone Tomorrow: On the Migration of MySpace Users

Mojtaba Torkjazi¹, Reza Rejaie¹, Walter Willinger²

¹ University of Oregon

² AT&T Labs-Research

- Provides explicit **profile status**
 - Public
 - Private
 - Invalid
- Availability of users' **last login**
 - Enables assessment of the level of activity among users
 - Importantly, allows inference of population growth of MySpace (see later for details)
- Global **visibility**
 - http://www.myspace.com/user_id
- **Monotonic** assignment of **numeric ID**

[Início](#)[Pessoas](#)[Encontrar amigos](#)[Música](#)[Vídeo ▼](#)[Jogos](#)[Mais ▼](#)

Tom

"keep up with what's hot on MySpace" <http://lnk.ms/6Y3xn>

Em 1 mar 2010

[exibir mais](#)

Exibir Meu/Minha: [Fotos](#) [Vídeos](#) [Blog](#) [Listas de Reprodução](#)

Tom Anderson
34 / Masculino
Los Angeles, California, US

- Feb. 26th 2009: MySpace ID space [1 ... 455,881,700]
- 50 parallel samplers to collect 360K users in less than 12 hours (0.1% of MySpace population)
- Using HTML parser to post-process the downloaded profiles and extract
 - User s' profile status (invalid, public, private)
 - Users' last login date
 - Users' friend list (only for public profiles)
- Unable to parse last login info for 0.96% of public and 0.08% of private profiles
 - Last login info is not provided or is provided with obvious errors (e.g. 1/1/0001)

- *Possible reasons behind MySpace's decline?*

- Slow-down in the growth rate of MySpace is related to **emergence of Facebook**
- Informal evidence (Alexa.com): Daily accesses to **Facebook** surpassed that of MySpace, at around **April 2008**



ACM SIGIR/SIGKDD WSDM 2011

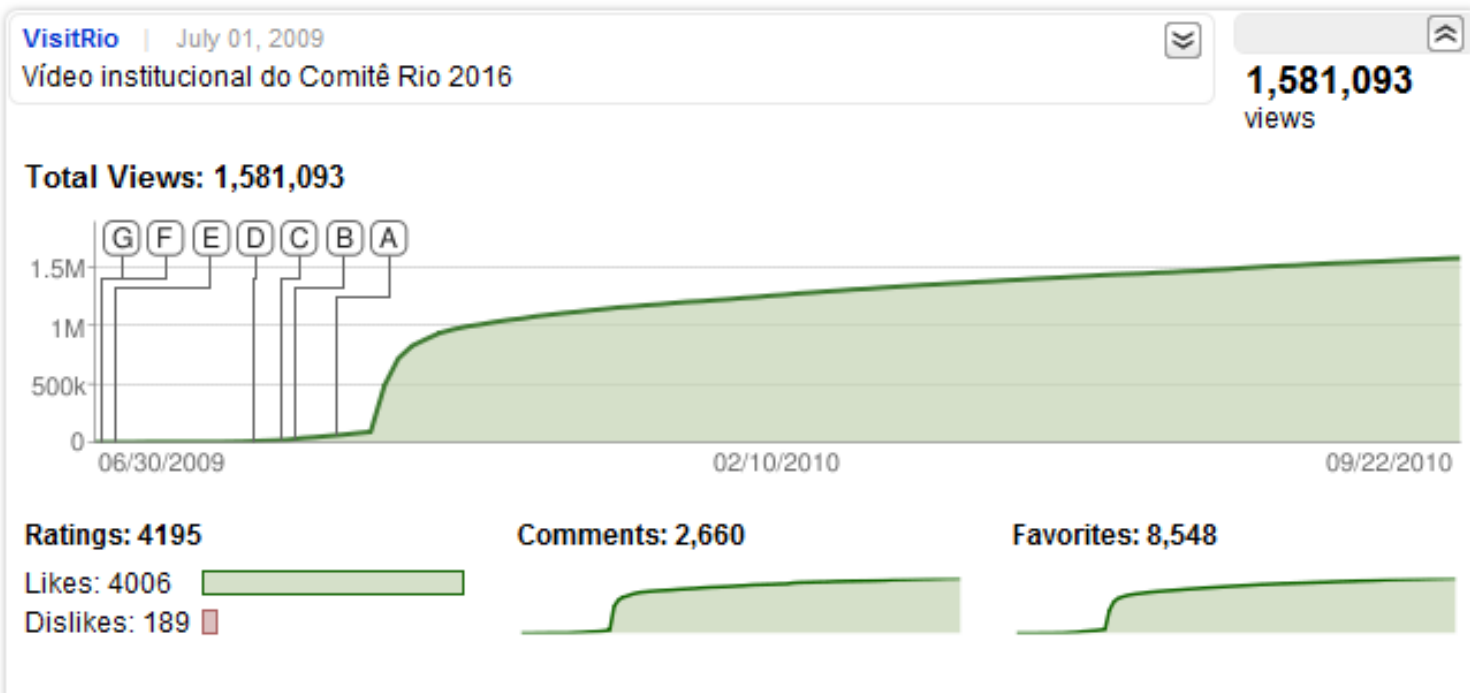
The Tube over Time: Characterizing Popularity Growth of YouTube Videos

F. Figueiredo¹, F. Benevenuto², J. Almeida¹

¹Universidade Federal de Minas Gerais (UFMG)

²Universidade Federal de Ouro Preto (UFOP)

Ajax no YouTube



Console HTML CSS Script DOM Net				
Clear Persist All HTML CSS JS XHR Images Flash				
URL	Status	Domain	Size	Timeline
GET watch_ajax?v=Z00jjc-WtZ	200 OK	youtube.com	11.6 KB	
http://chart.apis.google.com/chart?cht=lc&chs=625x120&chco=3d7930&chls=2,4,0&chg=0,-1,1,1&chxt=y,x&chxs=0,h&chxl=1:				
GET chart?cht=ls&chs=625x120	200 OK	chart.apis.google.com	1.3 KB	
GET chart?cht=ls&chs=625x120	200 OK	chart.apis.google.com	1.3 KB	
GET chart?cht=t&chs=625x120	200 OK	chart.apis.google.com	13.1 KB	
GET watch_ajax?v=Z00jjc-WtZ	200 OK	youtube.com	12.4 KB	
6 requests			49.4 KB	

Referrers no YouTube

Category	Referrer Type	Top			YouTomb			Random		
		t_{view}	f_{view}	f_{time}	t_{view}	f_{view}	f_{time}	t_{view}	f_{view}	f_{time}
EXTERNAL	First embedded view First embedded on First referrer from	0.57	0.11	0.35	0.81	0.16	0.41	0.07	0.08	0.22
FEATURED	First view from ad First featured video view	0.72	0.14	0.03	0.10	0.02	0.00	0.11	0.14	0.00
INTERNAL	First referrer from YouTube First referrer from Related Video	1.50	0.29	0.67	1.85	0.36	0.65	0.14	0.18	0.34
MOBILE	First view from a mobile device	0.26	0.05	0.51	0.02	0.00	0.02	0.03	0.03	0.05
SEARCH	First referrer from Google First referrer from YouTube search First referrer from Google Video	1.05	0.20	0.36	1.80	0.35	0.52	0.29	0.37	0.41
SOCIAL	First referrer from a subscriber First view on a channel page	0.36	0.07	0.35	0.01	0.00	0.01	0.01	0.00	0.12
VIRAL	Other / Viral	0.81	0.16	0.79	0.59	0.12	0.62	0.16	0.20	0.55

Table 5: Referrer categories and statistics (t_{view} : number of views (x 10^9); f_{view} : the fraction of views; f_{time} : fraction of times a referrer from the given category was the first referrer of a video).

WWW 2010

Earthquake Shakes Twitter User: Analyzing Tweets for Real-Time Event Detection

Takehi Sakaki
@tksakaki

Makoto Okazaki
@okazaki117
the University of Tokyo

Yutaka Matsuo
@ymatsuo

ACM IMC 2007

Measurement and Analysis of Online Social Networks

Alan Mislove, Massimiliano Marcon, Krishna Gummadi,
Peter Druschel, Bobby Bhattacharjee

Max Planck Institute for Software Systems (MPI-SWS)

This work

- Presents **large-scale measurement study and analysis** of the structure of multiple online social networks
 - 11 M users, 328 M links
- Data from four diverse online social networks
 - Flickr: photo sharing
 - LiveJournal: blogging site
 - Orkut: social networking site
 - YouTube: video sharing
- Our goals are two-fold:
 - Measure online social networks at scale
 - Understand static structural properties



Medição de OSNs



- Sites reluctant to give out data
 - Cannot enumerate user list
 - Instead, **performed crawls of user** graph
- Picked known seed user
 - Crawled all of his friends
 - Added new users to list
- Continued until all known users crawled
- Effectively **performed a BFS of graph**

Challenges faced

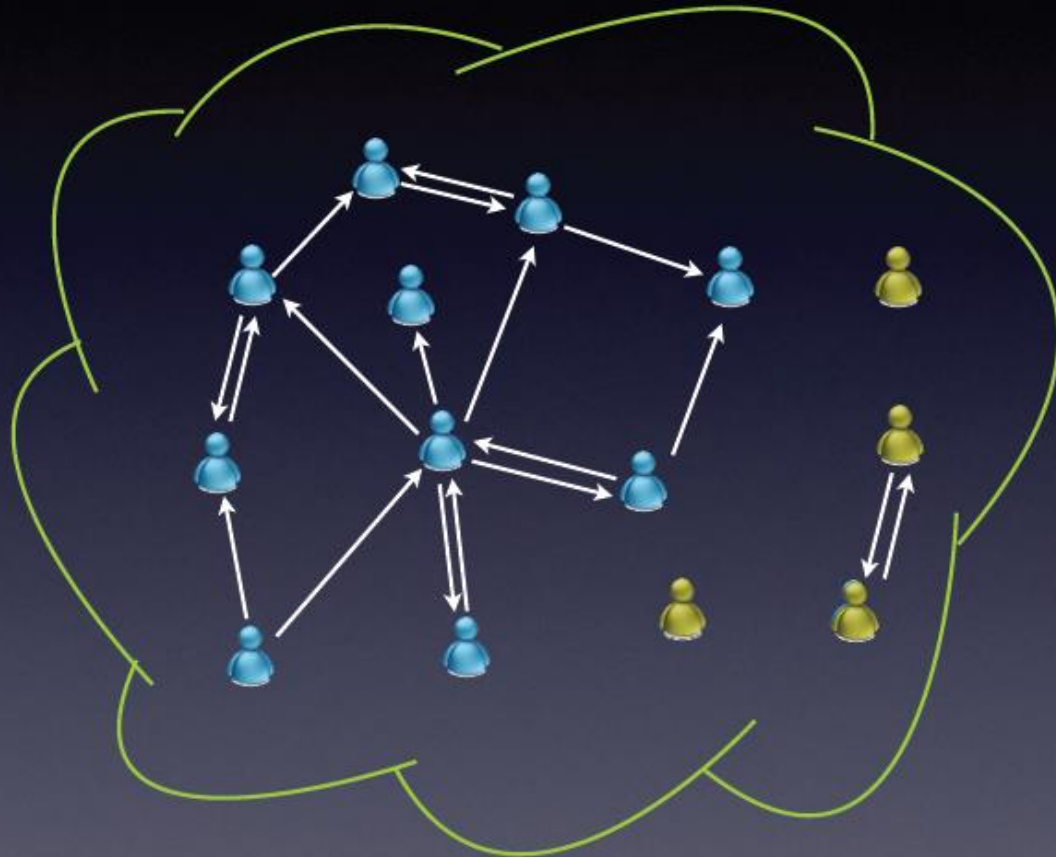
- Obtaining data using crawling presents unique challenges
- Crawling quickly
 - Underlying social networks changing rapidly
 - Consistent snapshot hard to get
 - Need to complete the crawl quickly
- Crawling completely
 - Social networks aren't necessarily connected
 - Some users have no links, or small clusters
 - Need to estimate the crawl coverage

How fast could we crawl?

- Crawled using cluster of 58 machines
 - Used APIs where available
 - Otherwise, used screen scraping
- **Crawls took varying times**
 - Flickr, YouTube: 1 day
 - LiveJournal: 3 days
 - Orkut (partial): 39 days
- Crawls **subject to rate-limiting**
 - Discovered appropriate rates



How much could we crawl?



- Users don't necessarily form single WCC
 - Disconnected users
- Estimate coverage by **selecting random users**
 - After crawl, determine fraction of users covered
- Networks tend to have **one giant WCC**

Evaluating coverage: Flickr



- Obtained random users by guessing usernames (#####@N00)
- Fraction of **disconnected users** is **73%**
- But, disconnected users have very low degree
 - 90% have no outgoing links, remaining 10% have few links
- Summary:
 - Covered 27% of user population, but remaining users have very few links

Evaluating coverage: LiveJournal

- Obtained random users using special URL
 - <http://www.livejournal.com/random.bml>
- Fraction of **disconnected users** is only 5%
- Summary:
 - Crawl covered 95% of user population

The LiveJournal logo, featuring the word "LIVEJOURNAL" in a bold, blue, sans-serif font, with a small trademark symbol (TM) to the right. The logo is set against a white rectangular background.

Evaluating coverage: Orkut



- At time of crawl, Orkut was fully connected
 - But, we ended crawl early
- How representative is our sub-crawl?
 - Performed multiple crawls from different seeds
 - Obtained random seed users using maximum-degree sampling
- **Properties consistent across smaller crawls**
- Summary:
 - Sub-crawl of user population, but likely representative of similarly sized subcrawls

Evaluating coverage: YouTube

- Could not obtain random users
 - Usernames user-specified strings
 - Not fully connected (could not use maximum-degree sampling)
- Unable to find estimate of user population
- Summary:
 - Unable to estimate fraction of users covered



Confirmou propriedades small-world

Network	C	Network	Avg. Path Len.
Web [2]	0.081	Web [12]	16.12
Flickr	0.313	Flickr	5.67
LiveJournal	0.330	LiveJournal	5.88
Orkut	0.171	Orkut	4.25
YouTube	0.136	YouTube	5.10

Redes sociais online possuem características Small World

Research on Online Social Networks: Time to Face the Real Challenges

Walter Willinger
AT&T Labs-Research
walter@research.att.com

Reza Rejaie, Mojtaba Torkjazi, Masoud Valafar
University of Oregon
{reza, moji, masoud}@cs.uoregon.edu

Mauro Maggioni
Duke University
mauro@math.duke.edu

ABSTRACT

Online Social Networks (OSNs) provide a unique opportunity for researchers to study how a combination of technological, economical, and social forces have been conspiring to provide a service that has attracted the largest user population in the history of the Internet. With more than half a billion of users and counting, OSNs have the potential to impact almost every aspect of networking, including measurements and performance modeling/analysis, network architecture and system design, and privacy and user behavior, to name just a few. However, much of the existing OSN research literature seems to have lost sight of this unique opportunity and has avoided dealing with the new challenges posed by OSNs. We argue in this position paper that it is high time for OSN researcher to exploit and face these opportunities and challenges to provide a basic understanding of the OSN eco-system as a whole. Such an understanding has to reflect the key role users play in this system and must focus on the system's dynamics, purpose and functionality

“large-scale and complex” network structures has relied on techniques from the tool box offered by “Network Science”. A hallmark of these techniques is that they tend to focus on graph metrics such as node degree distribution, clustering coefficient, density, diameter, or betweenness centrality that are purely descriptive in nature (e.g., see [3, 7] and also the discussion in [10]). As such, they say little or nothing about the graphs’ actual structure or dynamics. More importantly, they reduce OSNs to generic and static, and hence relatively uninteresting networked systems.

However, real-world OSNs are by nature highly dynamic structures. For example, in addition to the dynamics that is due to new users joining the system (generally by creating a new account) and existing users leaving the system (though typically without actively announcing their departure or closing their account), there is also the dynamics that results from active users interacting with each other. Here we focus on the former and note that at any point in time, the active users of an OSN may be just a fraction of

O que pode ser coletado

- Possibilidade de bloquear crawlers: **robots.txt**
 - Especifica diretórios e páginas que podem ou não podem ser coletadas com o uso de crawler

User-agent: Googlebot

Disallow: /confidencial

Disallow: /protegido

User-agent: *

Disallow: /temp

- Mais detalhes
 - <http://www.robotstxt.org/wc/robots.html>
 - <http://pt.wikipedia.org/wiki/Robots.txt>

Robots.txt – globo.com

User-agent: *

Disallow: /PPZ/

Disallow: /Portal/

Disallow: /Java/

Disallow: /Servlets/

Disallow: /GMC/foto/

Disallow: /FotoShow/

Disallow: /Esportes/foto/

Disallow: /Gente/foto/

Disallow: /Entretenimento/Ego/foto/

Disallow: /TVGlobo/CMA_Generico_Producao/tvg_repfoto_imagem_classe/

Robots.txt – orkut

```
User-agent: *  
Disallow: /Album.aspx  
Disallow: /AlbumZoom.aspx  
Disallow: /Block.aspx  
Disallow: /ClickTracker.aspx  
Disallow: /Community.aspx  
Disallow: /Communities.aspx  
Disallow: /CommEvent.aspx  
Disallow: /CommEvents.aspx  
Disallow: /CommMembers.aspx  
Disallow: /CommMsgs.aspx  
Disallow: /CommPolls.aspx  
Disallow: /CommPollResults.aspx  
Disallow: /CommPollVote.aspx  
Disallow: /CommTopics.aspx  
Disallow: /Event.aspx  
Disallow: /Events.aspx  
Disallow: /EventEdit.aspx  
Disallow: /EventGuests.aspx  
Disallow: /EventAlbums.aspx  
Disallow: /EventExternal.aspx  
Disallow: /EventGuestsExternal.aspx  
Disallow: /ExternalAlbum.aspx  
Disallow: /ExternalAlbumZoom.aspx  
Disallow: /ExternalHome.aspx  
Disallow: /FavoriteVideos.aspx  
Disallow: /FavoriteVideoView.aspx
```

Agregadores de tráfego

- Proxies: reconstrução de transações e sessões
 - YouTube Traffic Characterization: A view from the Edge. **IMC'07**
 - Understanding Online Social Networks Usage from a Network Perspective. **IMC'09**
- Agregadores de redes sociais
 - Characterizing User Behavior in Online Social Networks. **IMC'09**

ACM IMC 2007

YouTube Traffic Characterization: A View From the Edge

Phillipa Gill¹, Martin Arlitt²,
Zongpeng Li¹, Anirban Mahanti³

¹Dept. of Computer Science, University of Calgary, Canada

²Enterprise Systems & Software Lab, HP Labs, USA

³Dept. of Computer Science and Engineering, IIT Delhi, India

GET: /watch?v=wQVEPFzkhaM

OK (text/html)

GET: /vi/fNaYQ4kM4FE/2.jpg

OK (img/jpeg)





GET: swfobject.js

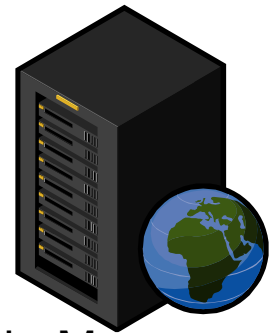
OK (application/x-javascript)

GET: /p.swf

OK (application/shockwave-flash)

GET: /get_video?video_id=wQVEPFzkhcM

OK (video/flv)



Campus

28.000 estudantes e 5.300 professores e funcionários

Link de 300Mb/s full-duplex

Objetivo:

Coletar o uso do YouTube em todo o campus

Obter dados de um período extenso

Proteger a privacidade dos usuários

Desafios:

Popularidade do YouTube

Limitação dos monitores de tráfego

Volume do uso da Internet do campus

- Identificar servidores provendo conteúdo do YouTube
- Utilizar **bro** para sumarizar cada transação HTTP em tempo real
- Reiniciar **bro** diariamente e comprimir o log diariamente
- Mapear cada visitante a um ID único

<http://www.bro-ids.org/>

Bro Intrusion Detection System



Version 1.0.3 - Last published Jun

Bro Overview

What is Bro?

Bro is an open-source, Unix-based Network Intrusion Detection System (NIDS) that passively monitors network traffic and looks for suspicious activity. Bro detects intrusions by first parsing network traffic to extract its application-level semantics and then executing event-oriented analyzers that compare the activity with patterns deemed troublesome. Its analysis includes detection of specific attacks (including those defined by signatures, but also those defined in terms of events) and unusual activities (e.g., certain hosts connecting to certain services, or patterns of failed connection attempts).

Bro uses a specialized policy language that allows a site to tailor Bro's operation, both as site policies evolve and as new attacks are discovered. If Bro detects something of interest, it can be instructed to either generate a log entry, alert the operator in real-time, execute an operating system command (e.g., to terminate a connection or block a malicious host on-the-fly). In addition, Bro's detailed log files can be particularly useful for forensics.

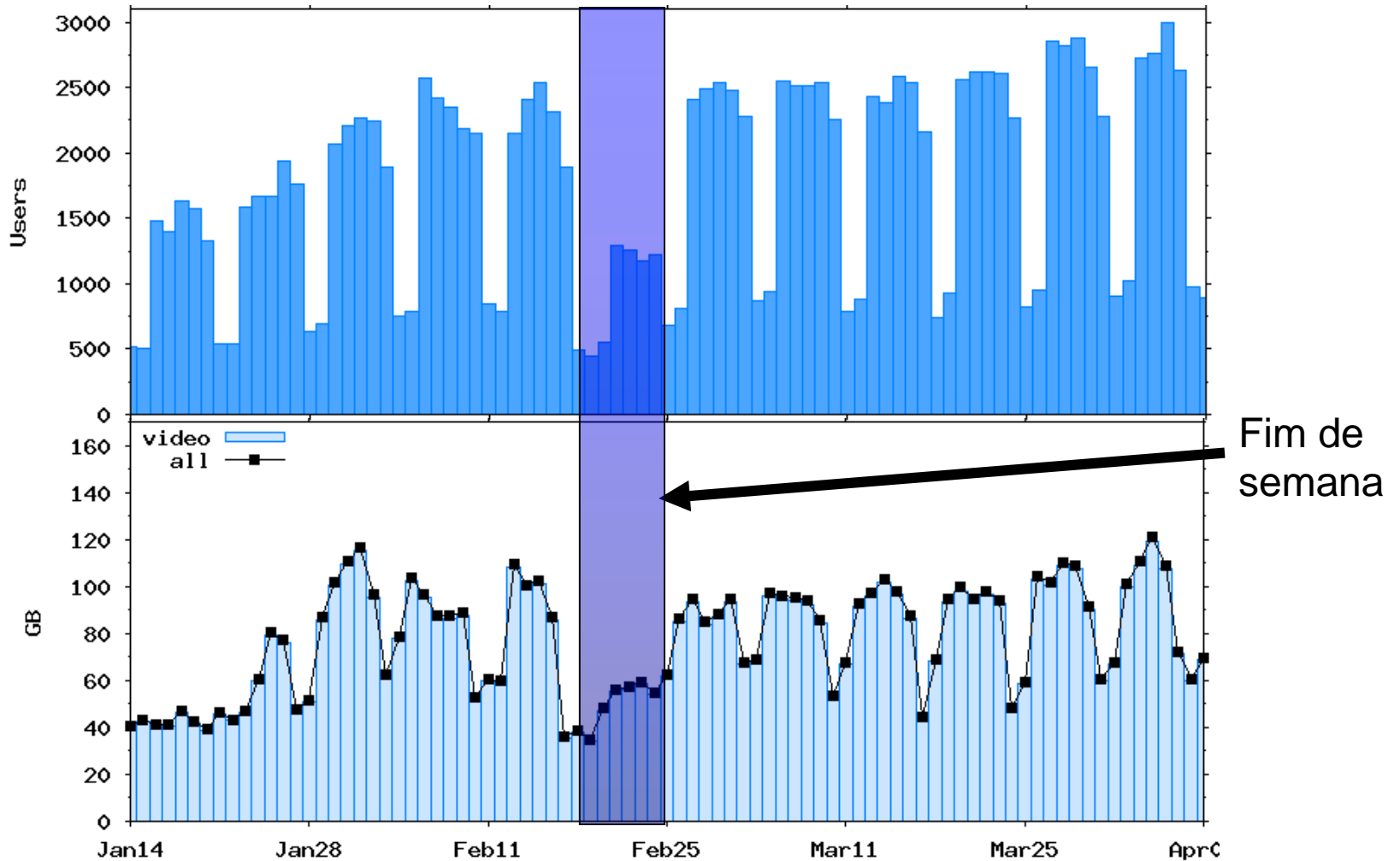
Bro targets high-speed (Gbps), high-volume intrusion detection. By judiciously leveraging packet-filtering techniques, Bro is able to achieve the necessary performance while running on commercially available PC hardware, and thus can serve as a cost-effective means of monitoring a site's Internet connection.

Start Date:	Jan. 14, 2007
End Date:	Apr. 8, 2007
Total Valid Transactions:	23,250,438
Total Bytes:	6.54 TB
Total Video Requests:	625,593
Total Video Bytes:	6.45 TB
Unique Video Requests:	323,677
Unique Video Bytes:	3.26 TB

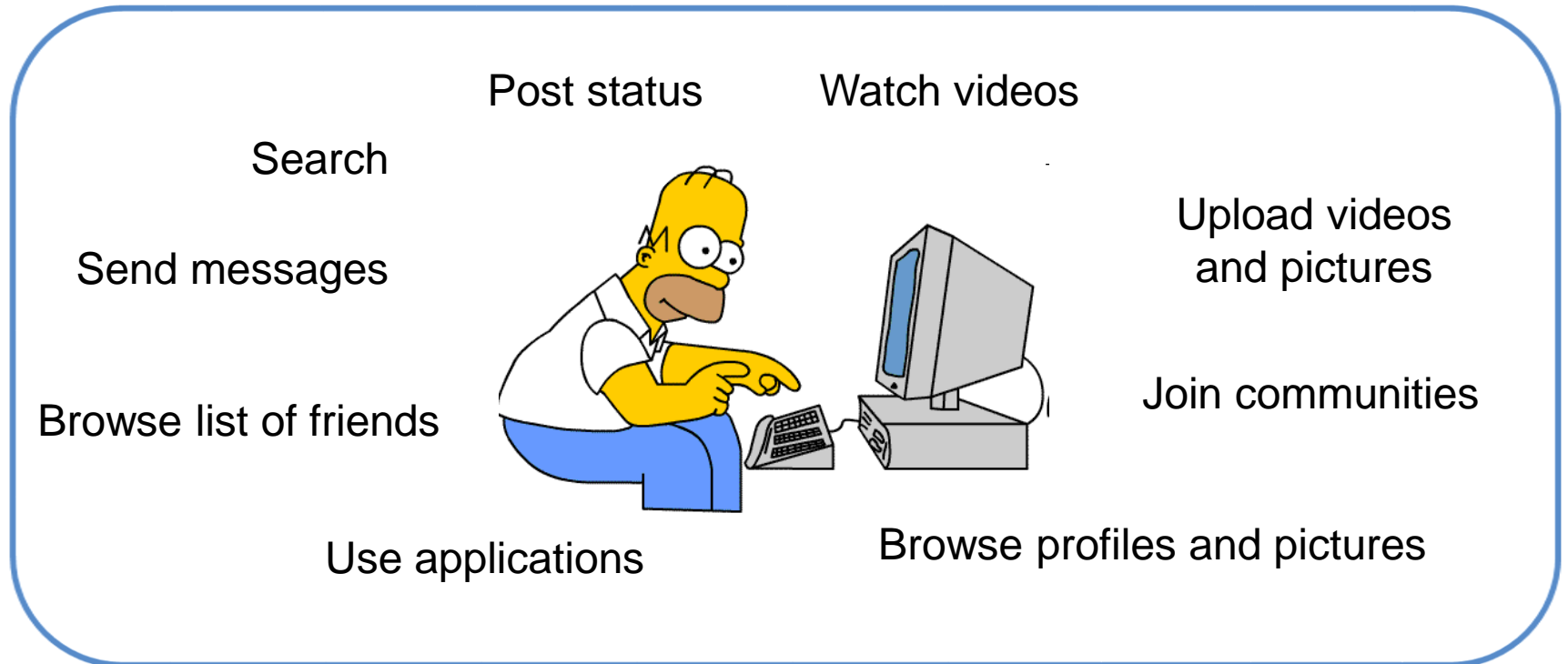
HTTP Response Codes

Code	% of Responses	% of Bytes
200 (OK)	75.80	89.78
206 (Partial Content)	1.29	10.22
302 (Found)	0.05	0.00
303 (See Other)	5.33	0.00
304 (Not Modified)	17.34	0.00
4xx (Client Error)	0.19	0.00
5xx (Server Error)	0.01	0.00

Campus Usage Patterns



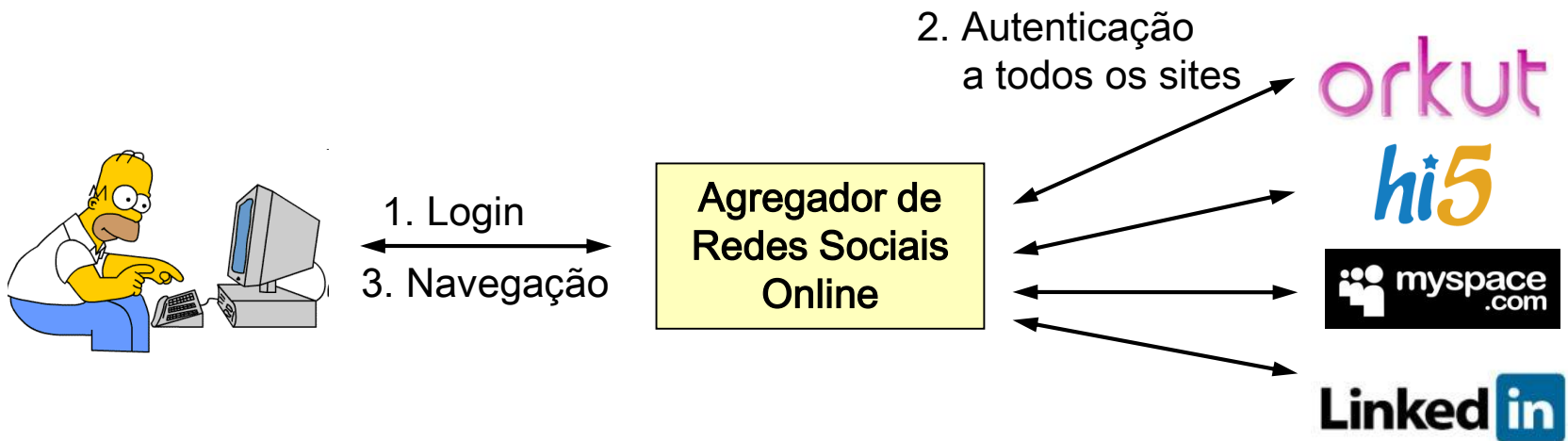
O que os usuários fazem nas redes sociais



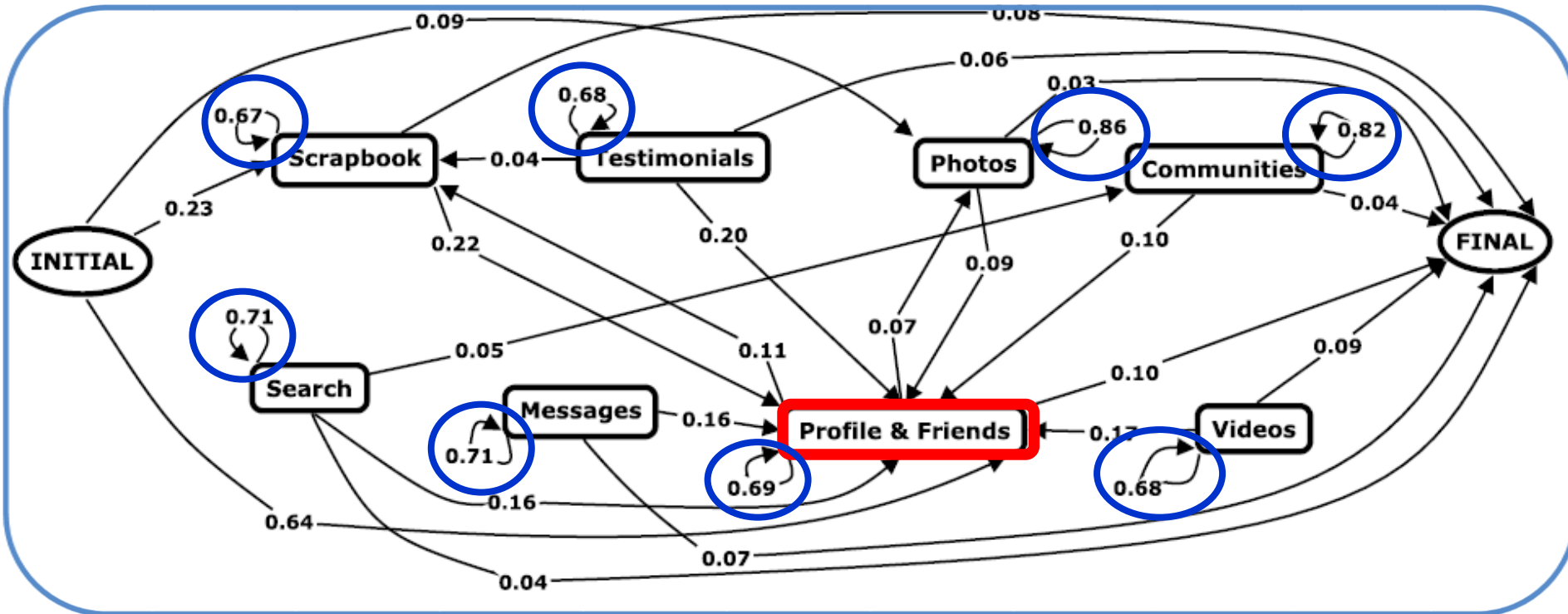
Entender navegação e interação dos usuários através de todas as atividades

Agregador de tráfego

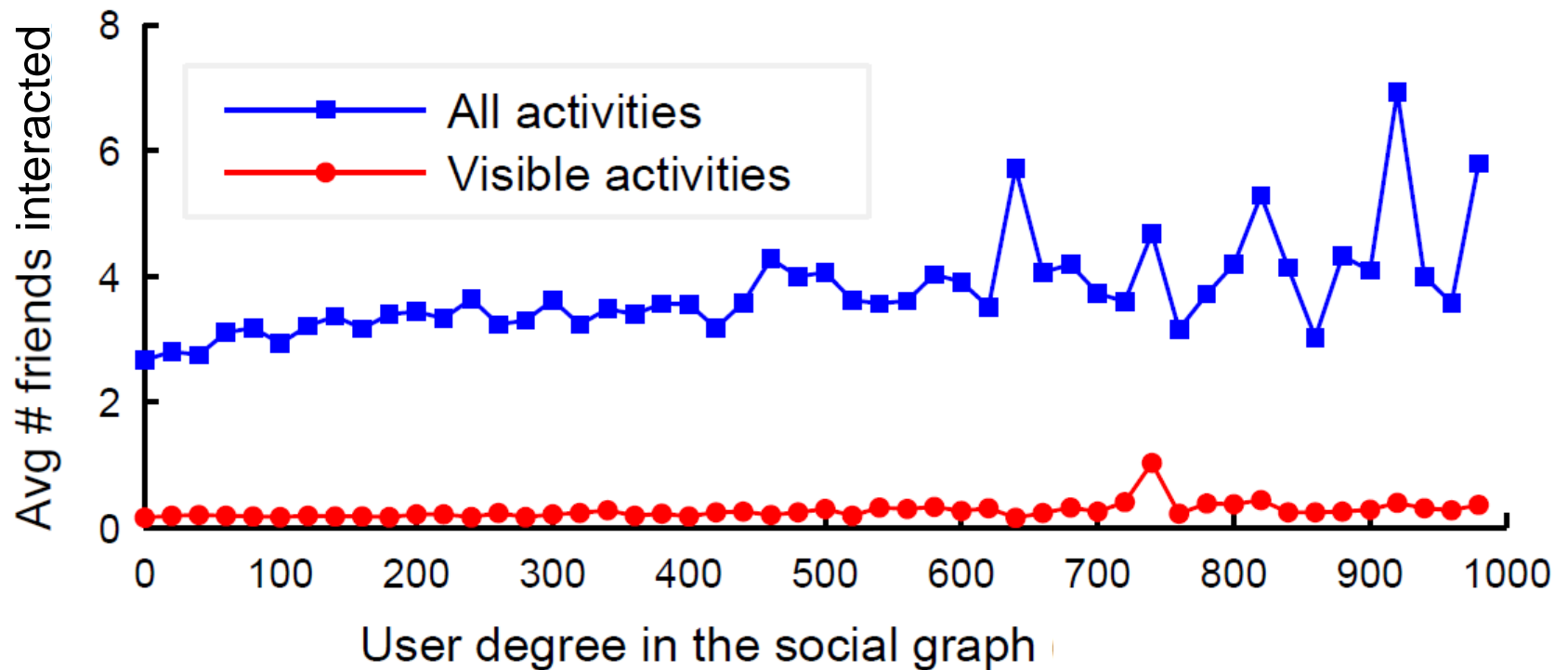
Dados podem ser coletados de um agregador de redes sociais



Seqüência das atividades



- Profile & Friends são centrais
- Self-loops são dominantes em todas as categorias



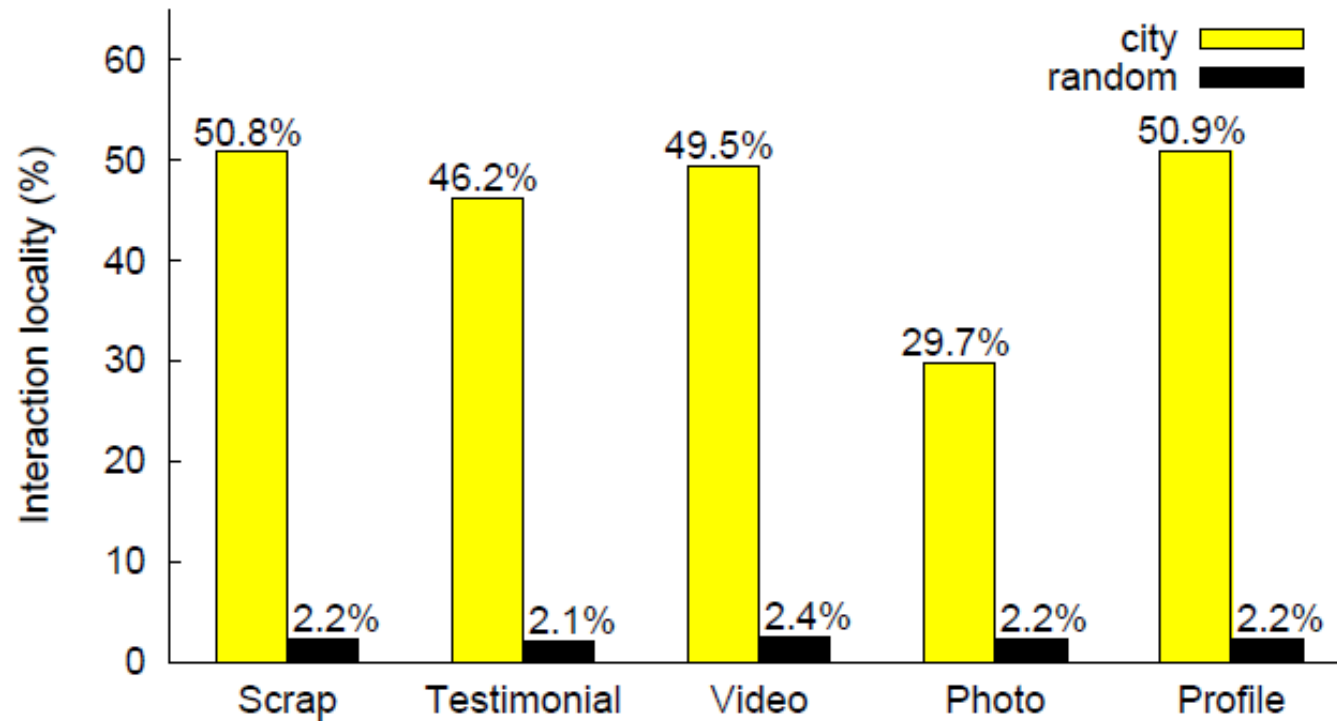
- Degree of interaction increases by an **order of magnitude** when incorporating silent interactions
- 85% of the active users showed **only silent interactions!**

- Informações geográficas são muitas vezes texto livre
 - Usuários podem preencher qualquer coisa. Ex. Sampa, BH, Marte
 - <http://developer.yahoo.com/maps/rest/V1/geocode.html>

Yahoo! Maps Web Services - Geocoding API

Finding Latitudes and Longitudes

The Geocoding Web Service allows you to find the specific latitude and longitude for an address. You can use this service to geocode your points in advance or forego it altogether with built-in geocoding in our AJAX and Flash APIs.



Conteúdo produzido e consumido localmente

ACM IMC 2009

Understanding Online Social Network Usage from a Network Perspective

Fabian Schneider¹, Anja Feldmann¹,
Balachander Krishnamurthy¹, Walter Willinger²

¹Technische Universität Berlin / Deutsche Telekom Laboratories

²AT&T Labs—Research

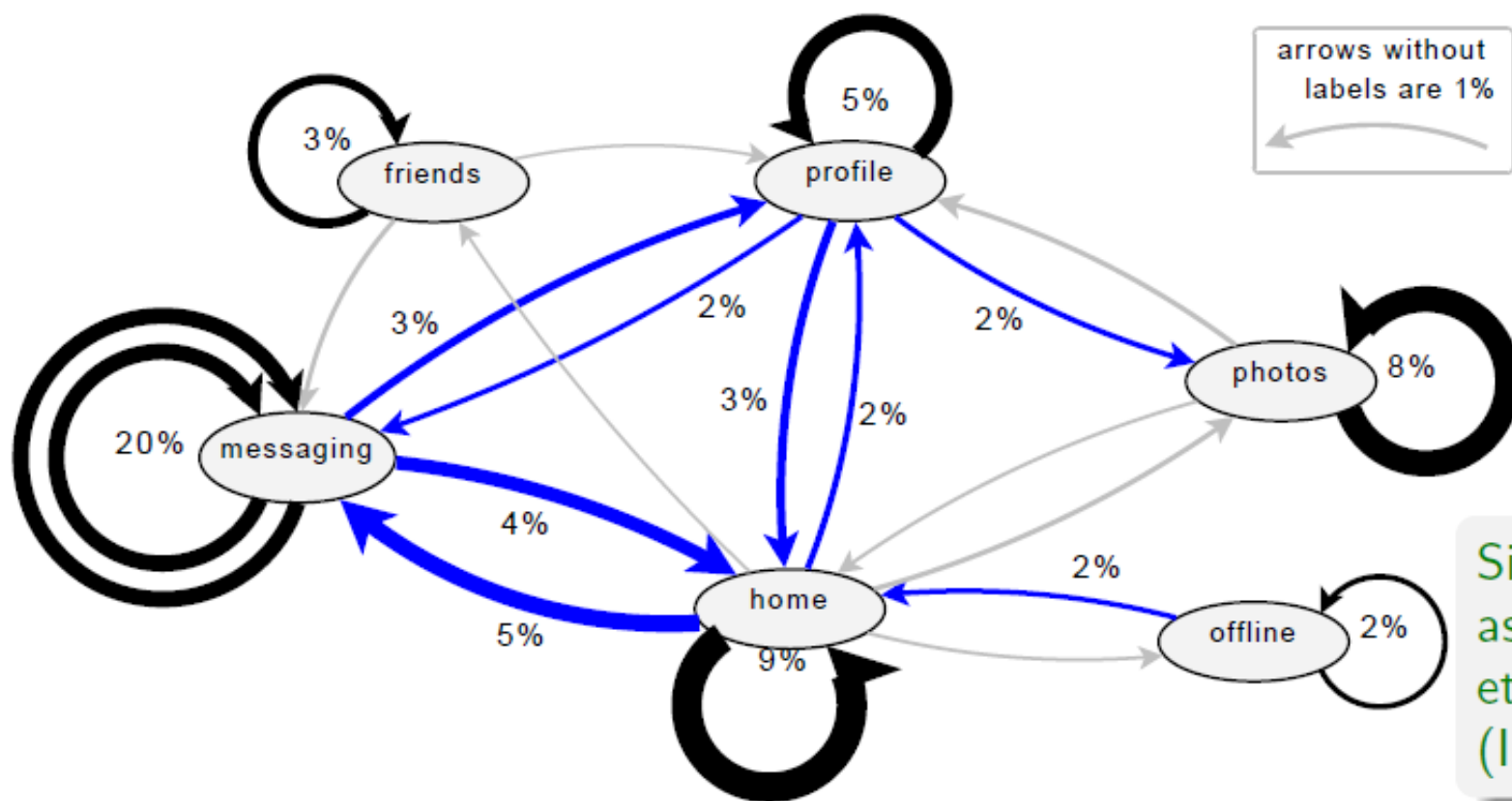
OSN Selection criteria:

- OSNs focussing on profiles (e. g., no YouTube, ...)
- 2 globally popular
- 2 locally popular (well represented at one ISP)



Seqüência de atividades

Click sequences of Facebook for ISP-A2: Global transition probabilities



Similar findings as Benevenuto et al for Orkut (IMC'09)

Findings

⇒ Messaging traps users; Home, Photos and Profile attract users to stay

Aplicações e jogos online

- Funcionamento e construção de aplicações em redes sociais
 - Unveiling Facebook: A measurement study of social network based applications. IMC'08
- Jogos Online
 - Social influence and the diffusion of user-created content. EC'09.

Aplicações

- Dominante em vários sistemas
 - Facebook, Orkut, Hi5, MySpace
- Duas plataformas maiores
 - Facebook Developer Platform (FDP)
 - OpenSocial

Facebook - aplicações

- Mais de 1 milhão de desenvolvedores em 180 países
- Mais de 550 mil aplicações ativas
- Mais de 100 milhões de usuários utilizando aplicações

Facebook - aplicações



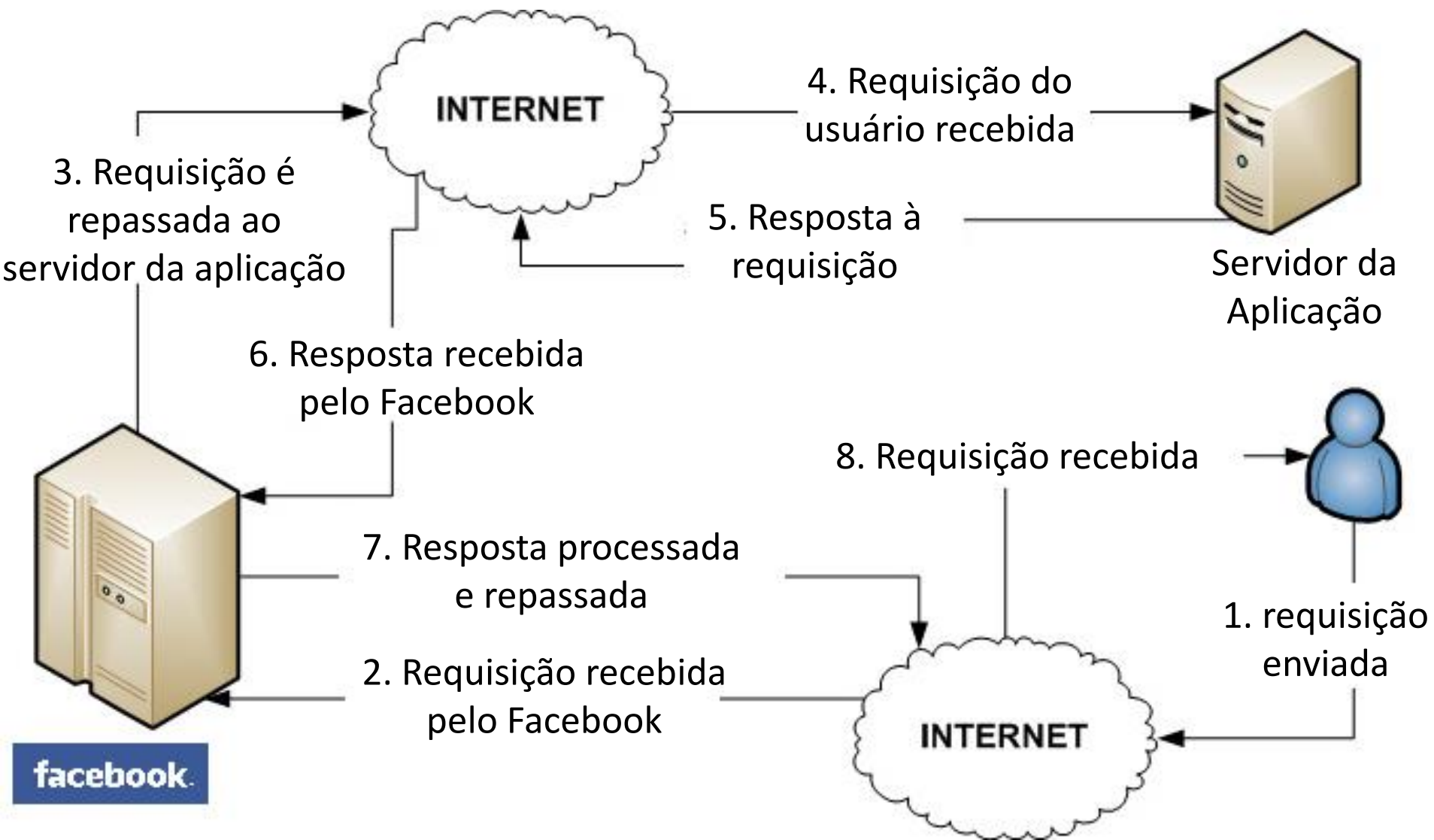
Facebook - Aplicações



Orkut - Aplicações



Aplicações



ACM SIGCOMM IMC 2008

Unveiling Facebook: A measurement
study of social network based Applications

A. Nazir, S. Raza, C. Chuah

University of California, Davis

Our Applications

- We deployed three applications on Facebook:



Fighters' Club (FC, 3.4M+, Jun 2007)

Social Gaming



Got Love? (GL, 4M+, Nov 2007)

Social Utility



Hugged (0.7M+, Feb 2008)

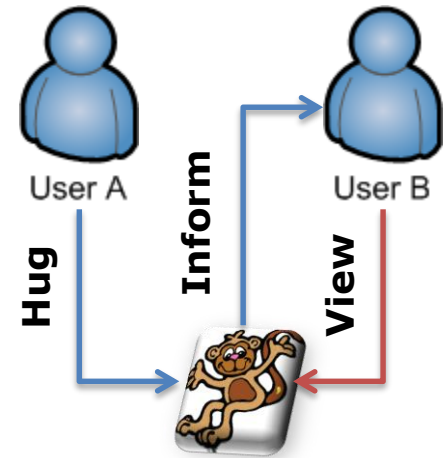
GL, HUGGED: SOCIAL UTILITY APPLICATIONS

♥ *GL*: friend-friend, one request per target friend

- *Hugged*: friend-friend, multiple requests per target friend

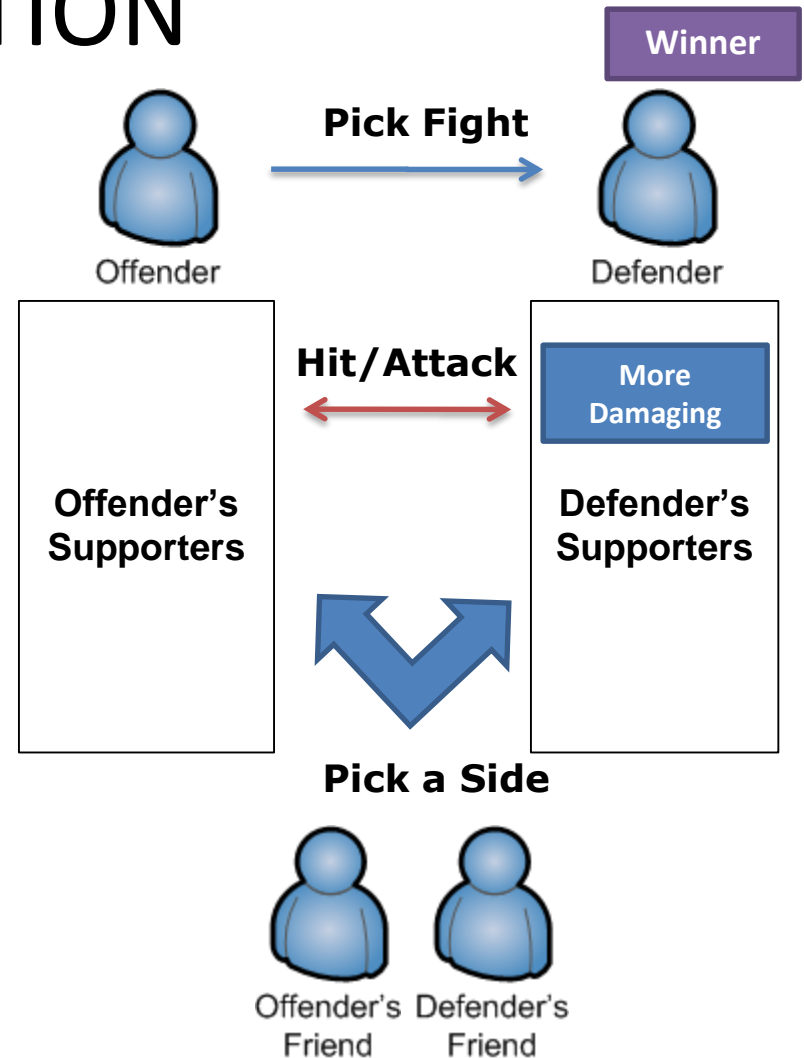


- Similar functionality:
 - User A hugs/loves (friend) User B
 - User B accepts/ignores hug/love



FIGHTERS' CLUB: A GAMING APPLICATION

- 🤪 Friend-friend, non-friend to non-friend interaction
- 🤪 Number of blows limited through points system



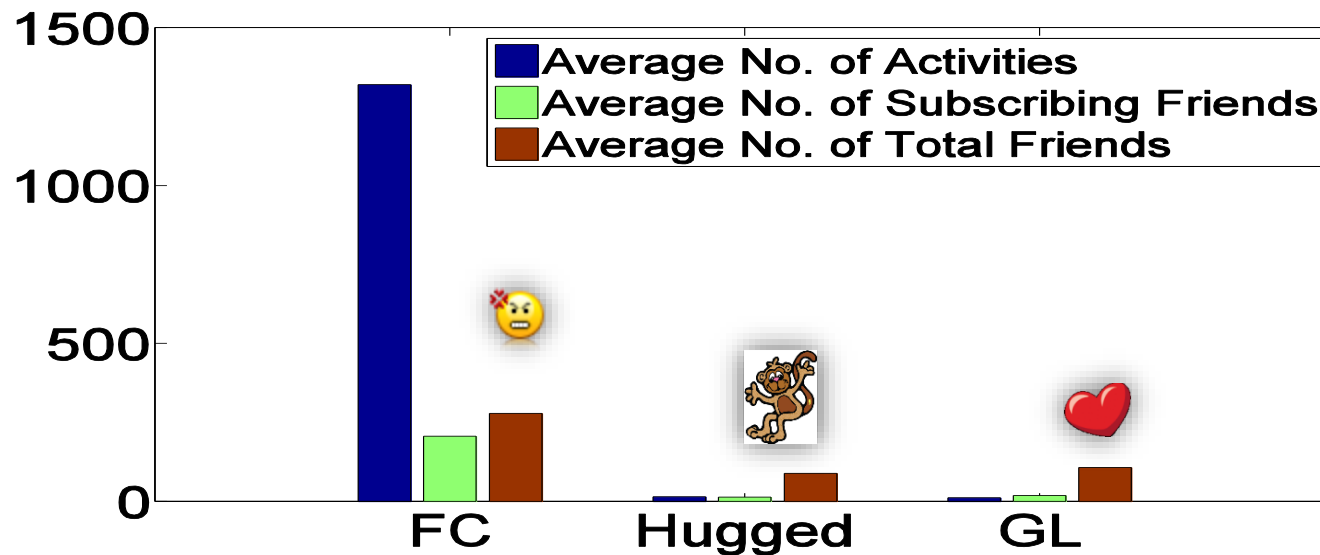
DATA SET SUMMARY

Table 1: Data set analyzed in this paper.

	Fighters' Club	Got Love	Hugged
Total Activities	25,911,335	7,196,302	2,146,819
Total Unique Users	154,681	5,376,704	1,322,631
Total Subscribing Users	85,928	1,518,767	408,651
Total Active Users	43,669	642,088	198,379
(Active) Users w/ Geo Info	40,982	97,465	180,216
Users w/ Friendship Data	35,349	72,074	121,389
BW Consumption Info	Dec 15 Onwards	Feb 15 Onwards	Feb 15 Onwards
Google Analytics Data	Dec 15 Onwards	Feb 15 Onwards	Mar 22 Onwards

SOCIAL GAMING VS. SOCIAL UTILITY APPLICATIONS: RESULTS

- Other differences:
 - Average number of activities higher on *FC* than on *GL*, *Hugged*
 - Average number of friends on application, total number of friends on Facebook, significantly higher for *FC* than *GL*, *Hugged*



INTERACTION GRAPHS:

DATA AND RESULTS SUMMARY

Table 3: Community Structures on Applications

	Fighters' Club	Got Love	Hugged
No. of Edges in Graph	16.8M	617,864	116,376
No. of Unique Users	73,300	277,540	51,343
Percentage of Users in Largest Component	91%	92.1%	86.7%
No. of Components	29	13,461	4,018
No. of Communities	51	1,951	521
Structure Coefficient	0.03	0.64	0.74
Max Size of Community	53,359	13,435	7,496
Max Geo Diversity	107	106	122
Max Network Diversity	2,858	2,285	1,084
Max Local in Community	2,852 (5.3%)	1,485 (34%)	455 (6.0%)
Clustering Coefficient	0.81	0.31	0.41
Diameter	10	45	29
Average Erdos-Renyi Clustering Coefficient	0.0062	0.000016	0.000085

INTERACTION GRAPHS: DATA AND RESULTS SUMMARY

Table 3: Community Structures on Applications

	Fighters' Club	Got Love	Hugged
No. of Edges in Graph	16.8M	617,864	116,376
No. of Unique Users	73,300	277,540	51,343
Percentage of Users in Largest Component	91%	92.1%	86.7%
No. of Components	29	13,461	4,018
No. of Communities	51	1,951	521
Structure Coefficient	0.03	0.64	0.74
Max Size of Community	53,359	13,435	7,496
Max Geo Diversity	107	106	122
Max Network Diversity	2,858	2,285	1,084
Max Local in Community	2,852 (5.3%)	1,485 (34%)	455 (6.0%)
Clustering Coefficient	0.81	0.31	0.41
Diameter	10	45	29
Average Erdos-Renyi Clustering Coefficient	0.0062	0.000016	0.000085

Actually Small World
Networks!

WWW 2008

Planetary-Scale Views on a Large Instant-Messaging Network

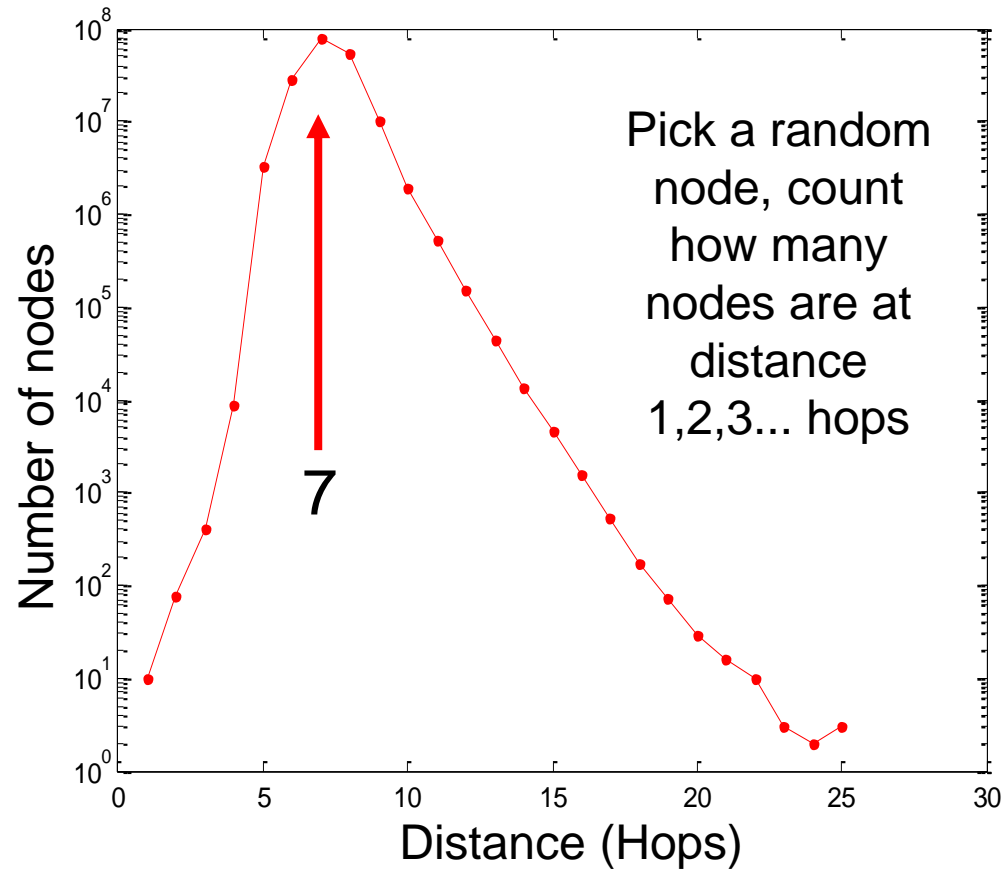
Jure Leskovec and Eric Horvitz

Carnegie Mellon University

Microsoft Research

Small-world effect

- Microsoft Messenger network
 - 180 million people
 - 1.3 billion edges
 - Edge if two people exchanged at least one message in one month period



WWW 2008

Comparison of Online Social Relations in Terms of Volume vs.
Interaction: A Case Study of Cyworld

Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue
Moon, Hawoong Jeong

KAIST

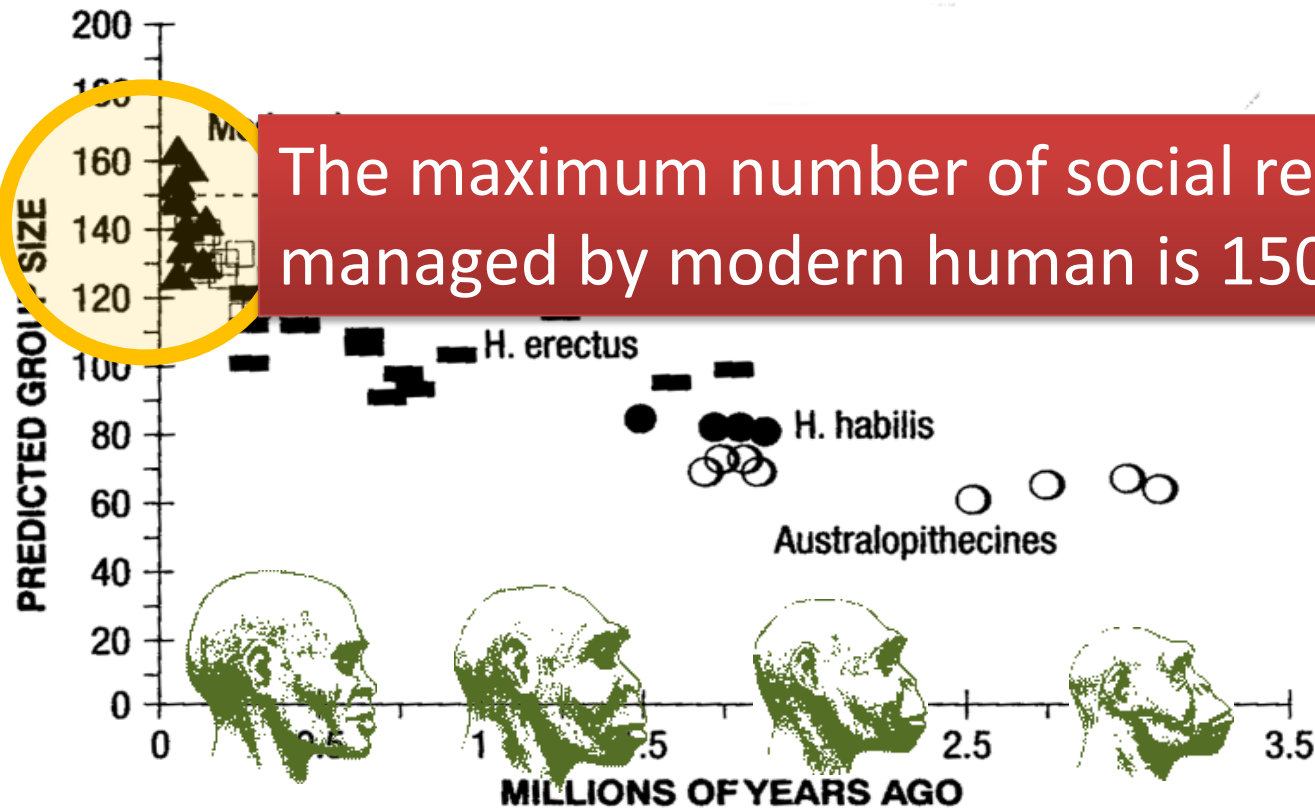
Cyworld



- Most popular OSN in Korea (22M users)
- Guestbook is the most popular feature
- Each guestbook message has 3 attributes
 - < From, To, When >
- We analyze 8 billion guestbook msgs of 2.5yrs

Dunbar's number

Behavioral and brain sciences, 16(4):681–735, 1993



The maximum number of social relations managed by modern human is 150.

Cyworld 200 vs. Dunbar's 150

- Has human networking capacity really grown?
 - Yes, technology helps users to manage relations
 - No, it is only an inflated number