Revisiting Preferential Attachment with applications to Twitter

Guillaume Ducoffe, Frédéric Giroire, Stéphane Pérennès, Stefano Ponziani

Université Côte d'Azur, Inria, CNRS, I3S, France

April $26^{\textbf{th}}$**, 2017**

Introducing myself

- **PHD in Computer Science** (Sept. 2014 – Dec. 2016)

"*Metric Properties of Large Graphs*"

under the guidance of David Coudert

team-project COATI (Université Côte d'Azur, Inria, CNRS, I3S, France)

- Research visits here and there

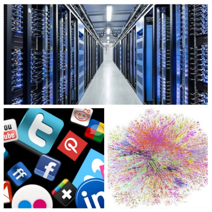
Columbia University, New York (with Prof. Chaintreau and Geambasu)

Universidad Adolfo Ibañez and **Universidad de Chile**, Santiago.

Some motivations for my research

Scalability in Network Algorithms

Growing size of communication networks



Social networks (Facebook $\geq$ 1.79 billion users)

Data Centers (Microsoft $\geq$ 1 million servers)

the Internet ($\geq$ 55811 Autonomous Systems)

"Efficient" algorithms on these graphs?

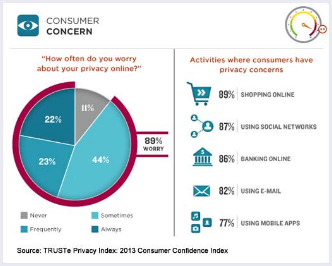
~~polynomial~~ $\rightarrow$ quasi-linear time
~~quadratic~~ $\rightarrow$ (sub)linear space

need for revisiting textbook (polynomial) graph algorithms

Some motivations for my research (cont'd)

Privacy in Network Algorithms

Raise of **privacy concerns** online



Online discrimination (Machine Learning, heuristics)

Violation of data policies (ex: Google App Education)

differential privacy: **preventing data leakage**

Web's transparency: **monitoring data use**

Research topics

Information propagation in networks $\Longrightarrow$ combinatorial problems on graphs

Finer-grained complexity analysis of graph problems

NP-hardness, **complexity in P**, *parallel complexity*, *query complexity*, ...

Metric tree-likeness in graphs

(with COATI team)

- Study of **geometric properties** of the (shortest) path distribution
- Computation of related parameters (**hyperbolicity**, **treelength**, **treebreadth**, treewidth)

algorithmic graph theory

Privacy at large scale in social graphs

(with Social Networks lab, Columbia)

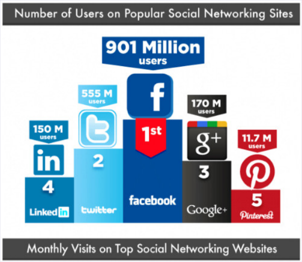
- Solution concepts for **dynamics of communities**
- Ad Targeting Identification

game and learning theory

Online Social Networks

Reasons for studying OSNs

Increasing social activity



(source: Go-Gulf.com, 2012)

Real-life applications:

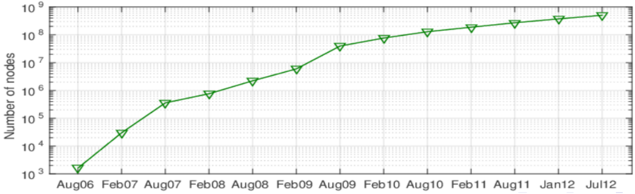
- sociology
- statistics
- **economy, advertising**
- **privacy**

Graph theoretical framework

In this talk: focus on **Twitter**

- $\sim 100M$ login/day
- in the Top 10 most visited websites
- 3^{rd} largest social media (?)

Twitter users between 2006 and 2012, by M. Gabielkov



Objectives

Design and **Analysis** of a **Random graph** model for Twitter

Some motivations:

- better knowledge of the structure
- predictive studies
- **Simulation** + **Testing** for algorithms

Related work: experiments on Twitter (1/2)

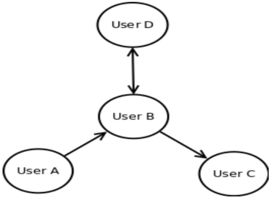
Conversation graph vs. **Graph of the followers**

[Cogan et al., Reconstruction and analysis of Twitter conversation graphs, '12]

In this talk: graph of the followers

Unidirectional relationships ("I'm interested in you")

- Follower: A follows B;
- Following: C is followed by B;
- Bidirectional: B and D follow each other.



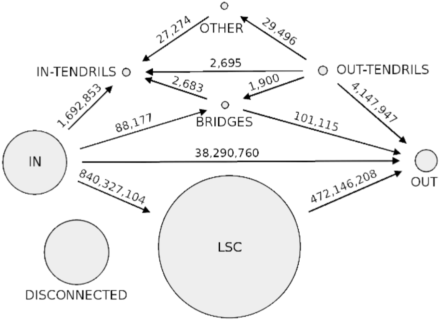
Related work: experiments on Twitter (2/2)

[Gabielkov et al.,'14]

- "Full" graph obtained by crawling

$\longrightarrow$ 505 million accounts interconnected by 23 billion links!

- "Macro structure" (dec. in strongly connected components)



LSC: 51% of users, 97% of following, 98% of followers.

Related work: undirected random model for networks

- **Erdös-Rényi**: "typical" graph

  each edge independently with probability p

Related work: undirected random model for networks

- **Erdös-Rényi**: “typical” graph

  each edge independently with probability p

- **Preferential attachment** paradigm: ”the rich gets richer”

- growing network (node + edge events)

- probability for a user to increase her degree is proportional to her current degree

[Barábasi-Albert, Bianconi-Barábasi, Watts-Strogatz, Chung-Lu, Krioukov et al., . . .]

Related work: undirected random model for networks

- **Erdös-Rényi**: “typical” graph

  each edge independently with probability p

- **Preferential attachment** paradigm: "the rich gets richer”

- growing network (node + edge events)

- probability for a user to increase her degree is proportional to her current degree

[Barábasi-Albert, Bianconi-Barábasi, Watts-Strogatz, Chung-Lu, Krioukov et al., ...]

Power-law:

$$Pr_v[deg(v) = k] = \Theta(k^{-a})$$

Related work: directed random model for networks

Few existing models and studies for digraphs

- **"directed Barábasi-Albert"** (node event + m outgoing arcs)
- **Bollobás et al.**: node events + 2 types of arc events (ingoing or outgoing arc)
  Remark: much more difficult to analyse!
- **RMAT [Chakrabarti et al., '04]**: fixed number of vertices and works with adjacency matrices

Our results

- An experimental study of the degree(s) distribution in the Twitter graph

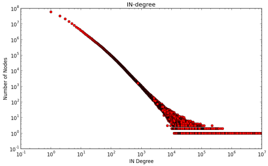
- Design of a new random digraph model

- Analysis of the model
  - experimental (comparisons with Twitter)
  - theoretical: **new techniques based on Markov processes**

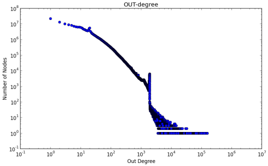
Experiments on the Twitter graph (1/4)

Degree(s) distribution in the LSC

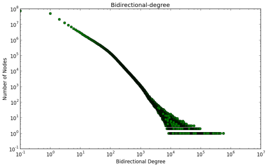
In-degree, Out-degree, Bidirectional follow Power-law distribution



in-degree



out-degree

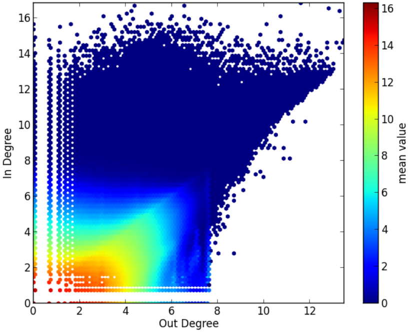


bidirectional

Experiments on the Twitter graph (2/4)

Linear correlations ? (Pearson's coefficient)

no OUT-IN correlation

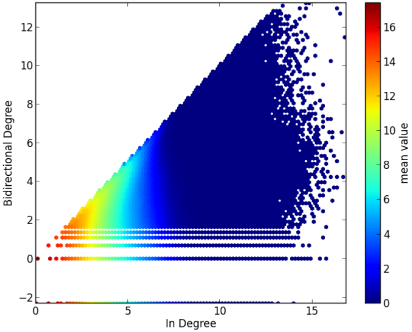


Pearson coefficient $\sim$ 0.1488

Experiments on the Twitter graph (3/4)

Linear correlations ? (Pearson's coefficient)

no IN-BI correlation

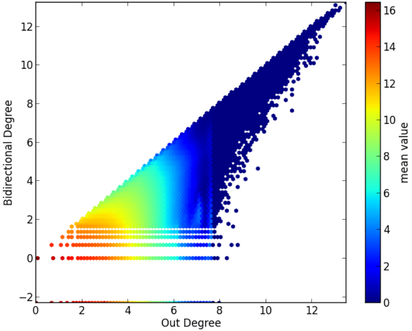


Pearson coefficient $\sim$ 0.1467

Experiments on the Twitter graph (4/4)

Linear correlations ? (Pearson's coefficient)

strong OUT-BI correlation



Pearson coefficient $\sim$ 0.9556

Limitations of existing models

Experiments vs. Bollobás et al. model

• The number of **bidirectional arcs** is high (theory predicts it should drop to zero)

• Strong positive correlation between out-degree and bidirectional degree (degrees should be "almost independent")

$\Longrightarrow$ **need for a new model that better accounts the specificities of Twitter**

Modelling: first attempt

Problem: number of bidirectional arcs is non vanishing (it should tend to zero)

Proposed solution: merge a **directed** random model with an **undirected** random model

$$\text{undirected edges} \longleftrightarrow \text{bidirectional arcs}$$

Issue: no correlation between out-degree and bidirectional degree !!

Modelling: second attempt

Modify **[Bollobás et al., '03]** for our needs.

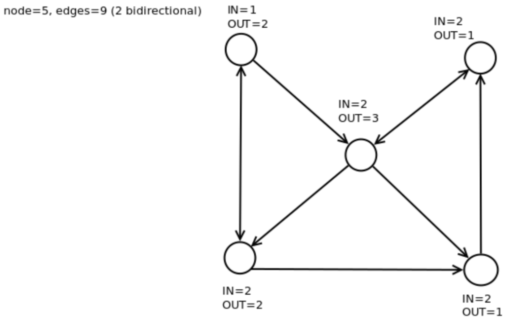
1) initial digraph $D(t_0)$;

2) **iterate**, for every time step $t \geq t_0$:

- addition of a new vertex with probability α (outgoing arc);
- addition of a new arc with probability $1 - \alpha$;
- **the new arc is bidirectional with probability** γ.

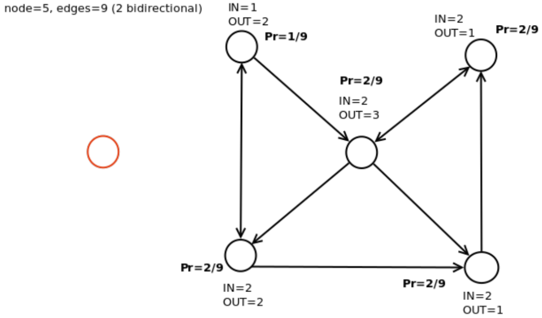
Examples

Initial digraph $D(t_0)$



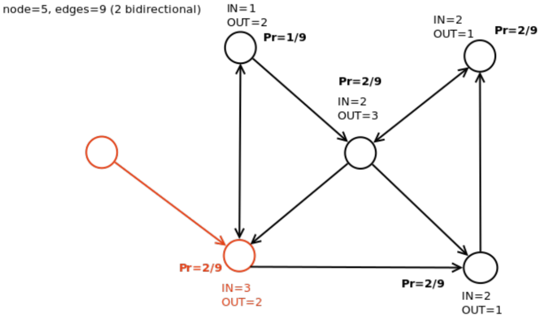
Examples

(A) **Node event**



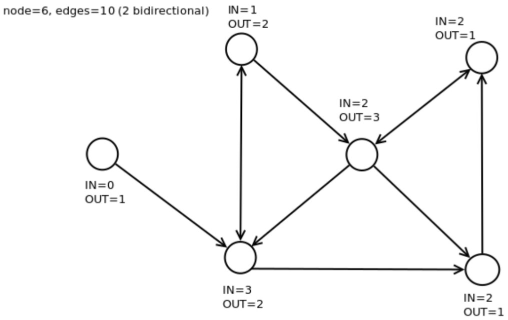
Examples

(A) **Node event**: add an out-going arc (with tail chosen w.r.t out-degree)



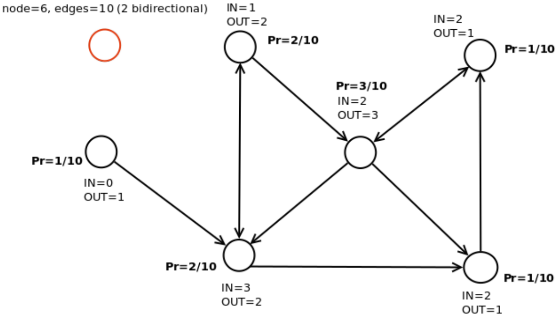
Examples

New digraph $D(t_0 + 1)$.



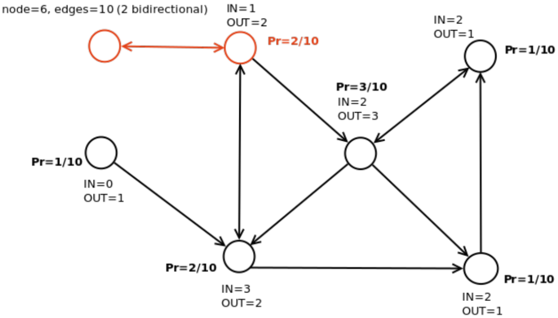
Examples

(B) **Node event**



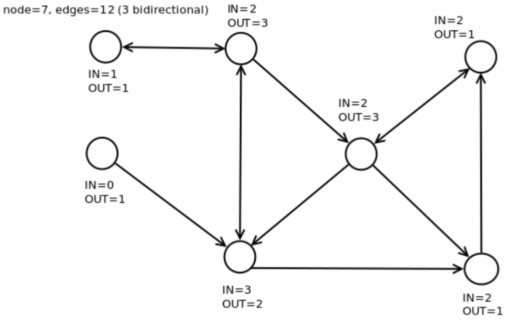
Examples

(B) **Node event**: add a bidirectional arc (with 2^{nd} end chosen w.r.t out-degree)



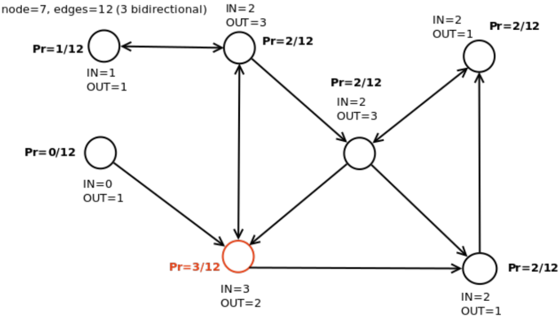
Examples

New digraph $D(t_0 + 2)$.



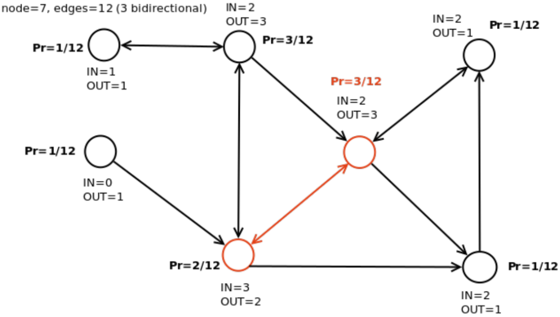
Examples

(C) **Arc event**: choose head w.r.t. in-degree



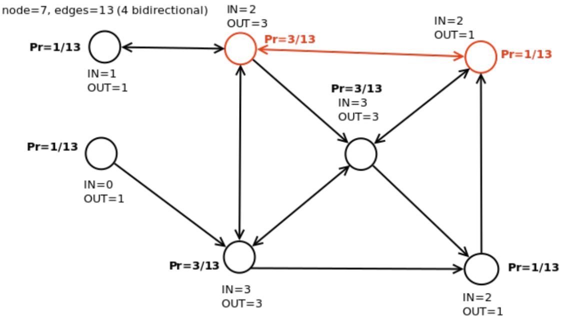
Examples

(C) **Arc event**: choose tail w.r.t. out-degree



Examples

(D) **Arc event**: choose ends w.r.t. out-degree



Degree Analysis

Computation of $x_{i,j,k}(t)$ = number of vertices, at the time step $t \geq t_0$, with:

in-degree $i + k$;

out-degree $j + k$;

bi-degree k.

Exact ? Asymptotic ?

Old school computations

borrow from **[Bollobás et al, '03]**.

recurrence equation:

$$\begin{aligned}
&\mathbb{E}[x_{i,j,k}(t+1) \mid D(t)] = x_{i,j,k}(t)\\
&+\frac{(1-\gamma)}{e(t)+\delta_{in}\cdot n(t)}\left((i+k-1+\delta_{in})\cdot x_{i-1,j,k}(t)-(i+k+\delta_{in})\cdot x_{i,j,k}(t)\right)\\
&+\frac{(1-\gamma)(1-\alpha)}{e(t)+\delta_{out}\cdot n(t)}\left((j+k-1+\delta_{out})\cdot x_{i,j-1,k}(t)-(j+k+\delta_{out})\cdot x_{i,j,k}(t)\right)\\
&+\frac{\gamma(2-\alpha)}{e(t)+\delta_{out}\cdot n(t)}\left((j+k-1+\delta_{out})\cdot x_{i,j-1,k}(t)-(j+k+\delta_{out})\cdot x_{i,j,k}(t)\right)
\end{aligned}$$

$$e(t) = t,\ \ n(t) = \Theta(t)\ (\mathit{Chernoff})$$

Old school computations (cont'd)

borrow from **[Bollobás et al, '03]**.

Case $i \to \infty,\ j, k$ fixed

by triple induction on i, j, k:

$$x_{i,j,k}(t)/t = \Theta_{j,k}\left(i^{-\left(1+\frac{1}{c_1}+(1+\delta_{out})(\frac{c_2+c_3}{c_1})\right)}\right)$$

Analysis fails in the other cases!

Relationship with Markov processes

$$
\begin{aligned}
&(t+1)\cdot \bar{x}_{i,j,k}(t+1) = t\cdot \bar{x}_{i,j,k}(t)\\
&+\frac{(1-\gamma)}{1+\alpha\delta_{in}}\left((i+k-1+\delta_{in})\cdot \bar{x}_{i-1,j,k}(t) - (i+k+\delta_{in})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+\frac{(1-\gamma)(1-\alpha)}{1+\alpha\delta_{out}}\left((j+k-1+\delta_{out})\cdot \bar{x}_{i,j-1,k}(t) - (j+k+\delta_{out})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+\frac{\gamma(2-\alpha)}{1+\alpha\delta_{out}}\left((j+k-1+\delta_{out})\cdot \bar{x}_{i,j-1,k}(t) - (j+k+\delta_{out})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+t^{-\mathcal{O}(1)}
\end{aligned}
$$

Relationship with Markov processes

$$
\begin{aligned}
&(t+1)\cdot \bar{x}_{i,j,k}(t+1) = t\cdot \bar{x}_{i,j,k}(t)\\
&+\frac{(1-\gamma)}{1+\alpha\delta_{in}}\left((i+k-1+\delta_{in})\cdot \bar{x}_{i-1,j,k}(t)-(i+k+\delta_{in})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+\frac{(1-\gamma)(1-\alpha)}{1+\alpha\delta_{out}}\left((j+k-1+\delta_{out})\cdot \bar{x}_{i,j-1,k}(t)-(j+k+\delta_{out})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+\frac{\gamma(2-\alpha)}{1+\alpha\delta_{out}}\left((j+k-1+\delta_{out})\cdot \bar{x}_{i,j-1,k}(t)-(j+k+\delta_{out})\cdot \bar{x}_{i,j,k}(t)\right)\\
&+t^{-\mathcal{O}(1)}
\end{aligned}
$$

transitions

- $(i,j,k)\to(i+1,j,k)$ with rate $\frac{(1-\gamma)}{1+\alpha\delta_{in}}(i+k+\delta_{in})$
- $(i,j,k)\to(i,j+1,k)$ with rate $\frac{(1-\gamma)(1-\alpha)}{1+\alpha\delta_{out}}(j+k+\delta_{out})$
- $(i,j,k)\to(i,j,k+1)$ with rate $\frac{\gamma(2-\alpha)}{1+\alpha\delta_{out}}(j+k+\delta_{out})$

rebirth process (new nodes)

Main tool

$\overrightarrow{X}(t) = (x_{i,j,k}(t))_{i,j,k}$
Q **rate matrix**

$$(t+1)\cdot[\overrightarrow{X}(t+1) - \overrightarrow{X}(t)] = Q \cdot \overrightarrow{X}(t) + \overrightarrow{o}(t^{-\mathcal{O}(1)})$$

Theorem

If the Markov process admits a stationary distribution Π then a.a.s. $\overrightarrow{X}(t) \to \Pi$

Sketch of proof

$$(t+1)\cdot[\overrightarrow{X}(t+1)-\overrightarrow{X}(t)] = Q\cdot\overrightarrow{X}(t)+\overrightarrow{o}(t^{-\mathcal{O}(1)})$$

continuum theory [Barabási-Bianconi,'00]

Reinterpret:

$$[\overrightarrow{X}(t+1)-\overrightarrow{X}(t)] = \frac{1}{(t+1)-t}\cdot[\overrightarrow{X}(t+1)-\overrightarrow{X}(t)]$$

As:

$$\frac{d(\overrightarrow{X}(t))}{dt}$$

Sketch of proof (cont'd)

$$(t+1) \cdot \frac{d(\overrightarrow{X}(t))}{dt} = Q \cdot \overrightarrow{X}(t)$$

$P_Q(t) = Pr[$ in state (i, j, k) at time $t]$

$$\begin{cases} \frac{d(P_Q(t)}{dt} = Q \cdot P_Q(t) \\ P_Q(t) \to \Pi \end{cases}$$

$$\overrightarrow{X}(\mathbf{t}) = \mathbf{P_Q}(\ln(\mathbf{t}+\mathbf{1})) \to \mathbf{\Pi}$$

Applications to Preferential attachment models

- 1-dimensional (**undirected graphs**)

Ex: Chung-Lu model.

$$(f(i+1)+1)S_{i+1} = f(i)S(i)$$

- Existence of stationnary distribution: $\sum_j \frac{1}{f(j)}$ diverges;

- if $F(i) = \int^i dt/f(t)$ then:

$$S_i = \Theta(\exp[-F(i)]/f(i))$$

- 2-dimensional (**directed graphs**) $\longrightarrow$ exact closed-form formula

Bollobás et al.

- 3-dimensional (**our case**): reduction to 2-dimensional case

Conclusion

- First study of the degree(s) distribution on Twitter
- Design and Analysis of a new random digraph model
- Automation through **Markov processes**

Perspectives

On-going work !

- New applications of our approach ?

- Extend study to other properties of Twitter ?

Mersi!