Paper Reading

Quick Detection of High-degree Entities in Large Directed Networks

K. Avrachenkov
Inria
k.avrachenkov@inria.fr

N. Litvak
University of Twente
n.litvak@utwente.nl

L. Ostroumova Prokhorenkova
Yandex
ostroumova-la@yandex.ru

E. Suyargulova
Yandex
siyargul@yandex.ua

(from ICDM '14)

Jan 2015

Speaker: Kazuhiro Inaba

内容

- Social graph から
- Top-k high in-degree nodes を
- **Sublinear time で**
- 見つける

より具体的な問題設定 (1)

- Twitter の人気ユーザを知りたい
- しかし Twitter API には回数制限が！
- **頂点数≃1.5G より遙かに少ない API 呼出で**、人気ユーザを割り出したい

より具体的な問題設定 (2)

- できること
  - ユーザを uniformly random に選ぶ
  - API呼出1回で、ユーザ1人の入次数を得る
  - API呼出1回で、ユーザ1人の出辺を5000本まで得る

    (※仮定: 5000 ≫ ほとんどのユーザの出次数)

提案手法

（n_1+n_2回のAPI呼出でn_2個の<頂点,次数>対を返す）

```
for w in W do
    S[w] ← 0;
for i ← 1 to n_1 do
    v ← random(N);
    foreach w in OutNeighbors(v) ⊂ W do
        S[w] ← S[w] + 1;
w_1, ..., w_{n_2} ← Top_n_2(S) // S[w_1], ..., S[w_{n_2}] are
the top n_2 maximum values in S;
for i ← 1 to n_2 do
    d_i ← InDegree(w_i);
```

n_1 個の頂点を
ランダムに選ぶ

そこから多く指されている
n_2 個の頂点について

全体からの入次数を取得。
(Top-K が欲しければここからK個取る)

実験

- 実際に Twitter に対して
  $n_1 + n_2 = 1000$ 回APIを呼んで実験

- “Ground Truth” との近さで評価
  - ***Fraction***: 真の Top-K のうち何割を見つけられたか
  - ***First-Error Index***: 見つけたn_2個から漏れた最上位

実験: “Ground Truth”

- Twitter社は公開していない
- twittercounter.com という第三者サービスを参照しようとした

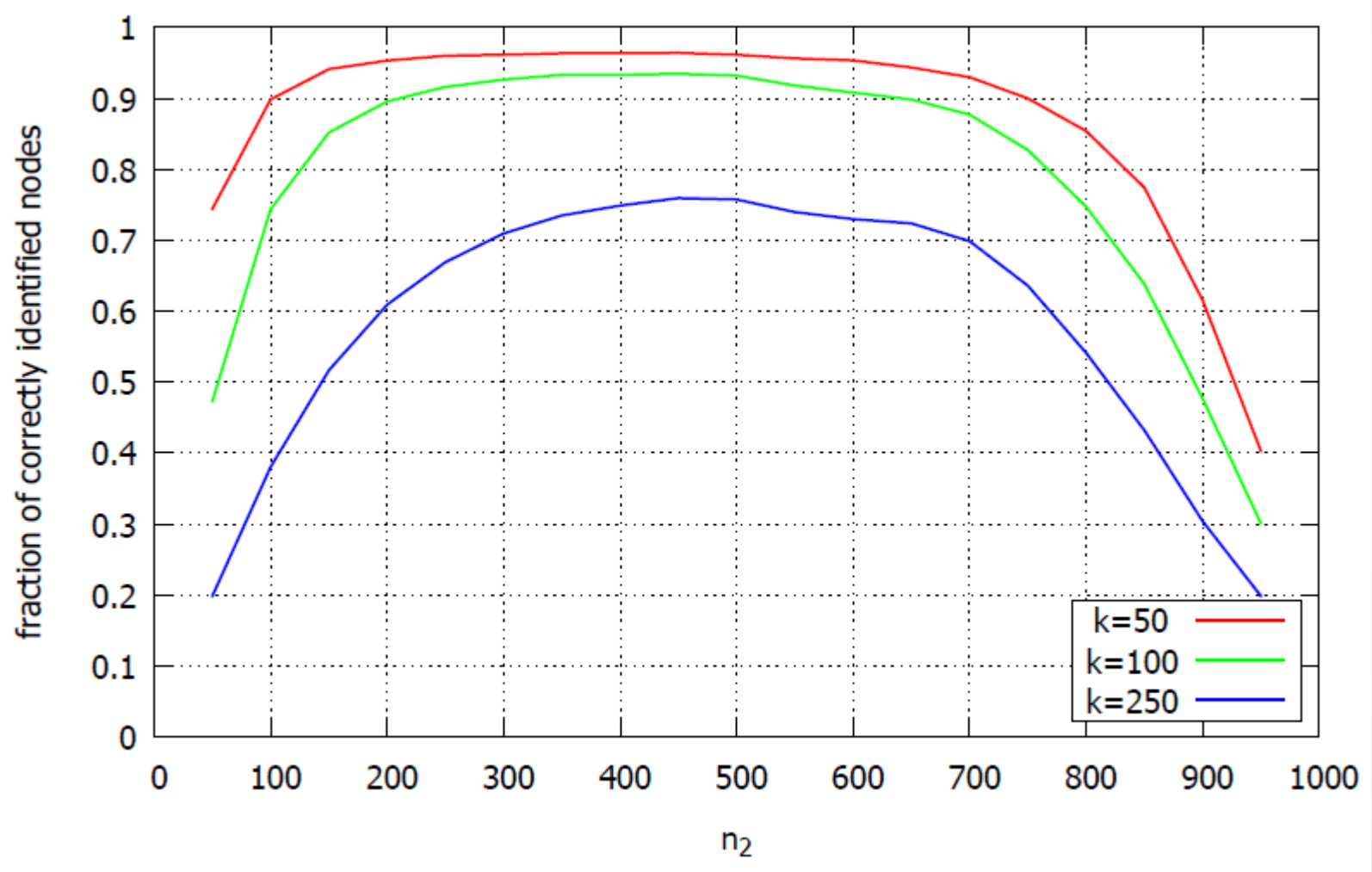
“Ground Truth”

- Twitter社は公開していない
- twittercounter.com という第三者サービスを参照しようとした

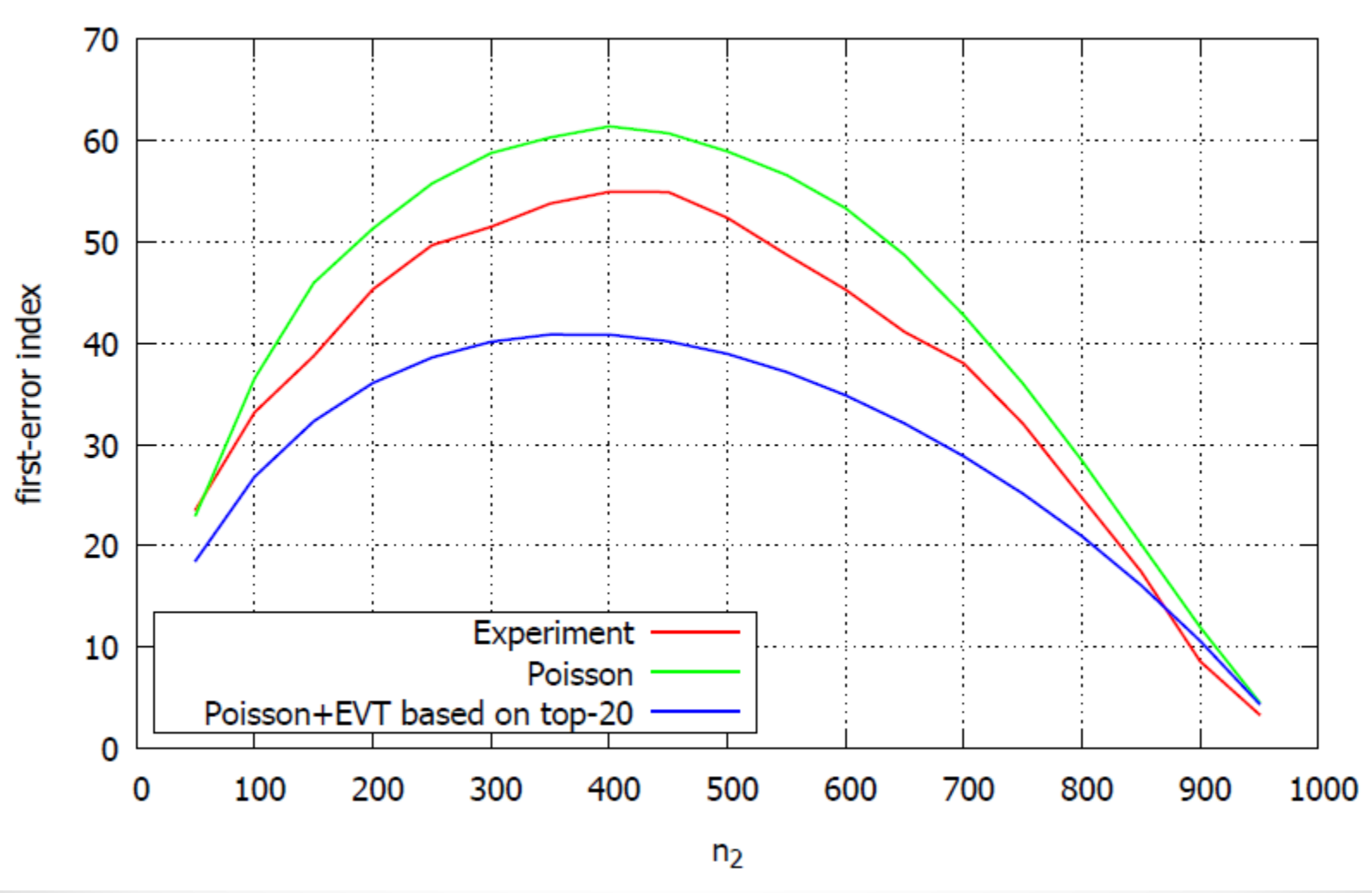
！！！しかし！！！

- 提案手法 [n_1=n_2=20,000] の方が精度がよかった
  ➔ [n_1=n_2=500,000] をGround Truthとして使用

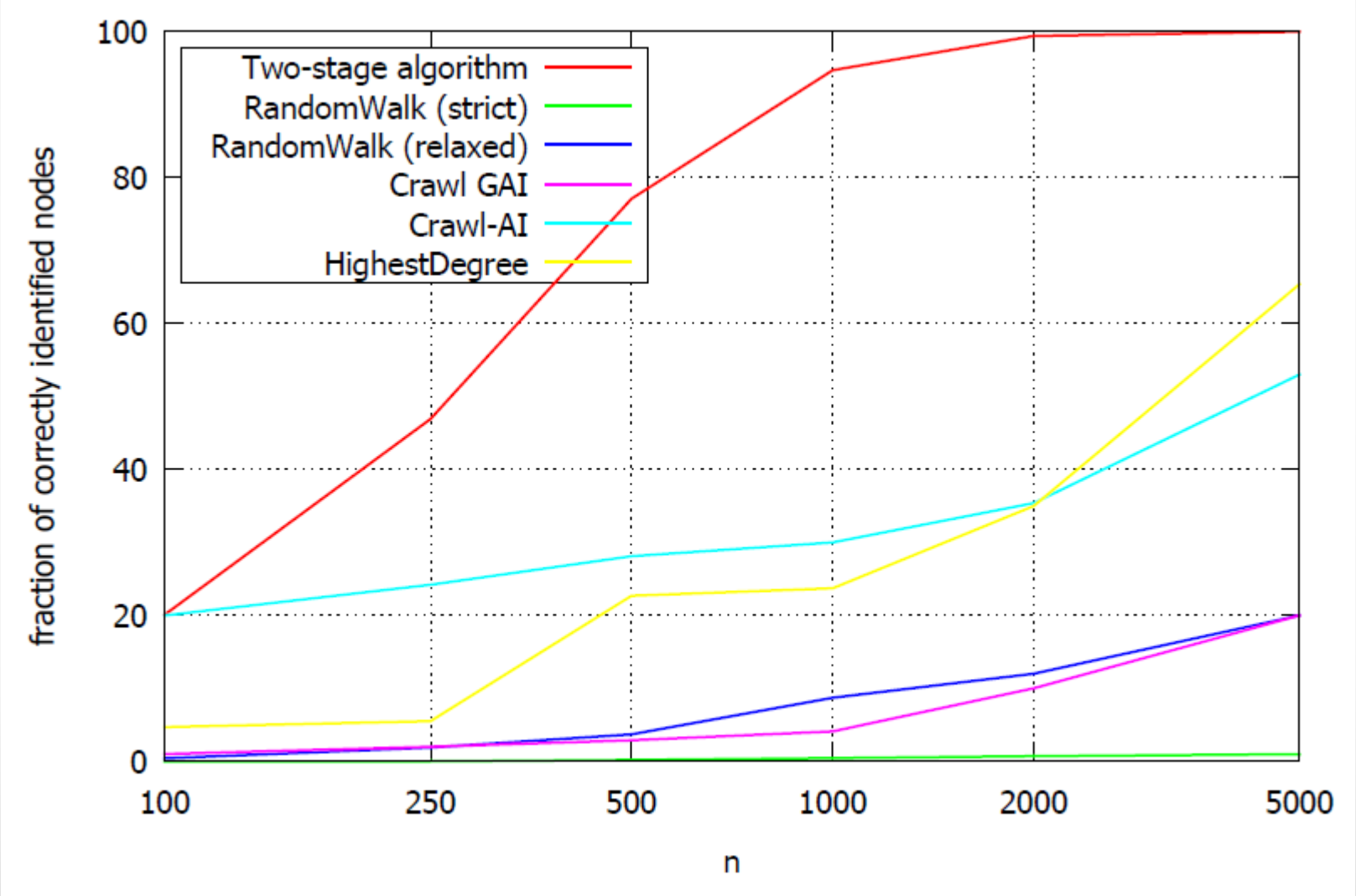
Twitterでの実験結果 (Fraction)



Twitterでの実験結果 (First-Error)



他の手法との比較 (Top-100; Fraction)



その他の応用

- 提案手法は「フォローする側」と「される側」が分かれている場合でもそのまま適用可能
  - 例：SNSで人気のある “グループ” は？
    - ロシアのSNS VK.com (～200M users) で実験。(n_1=700, n_2=300) で top-100 の73.2%を発見

解析

- n_1 と n_2 の最適なバランスは？
- $n = n_1 + n_2$ をどのくらい増やせば十分な精度が得られる？

n_1 と n_2 の最適なバランスは？

Proposition 1. *It is optimal to choose n_1 such that $n = O(n_1)$.*

= “n_1 に比べてあまり n_2 を大きくしても意味がない”

証明: そもそもfirst stageで
平均次数×n_1 頂点くらいしか選ばれないので。

どのくらい増やせば十分な精度？

Proposition 2. *For large enough n_1, the inequality*

$$Z_k(n_1) := \sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} \geqslant z_{1-\varepsilon}, \qquad (14)$$

where $z_{1-\varepsilon}$ is the $(1-\varepsilon)$-quantile of a standard normal distribution, guarantees that the mean fraction of top-k nodes in W identified by Algorithm 1 is at least $1-\varepsilon$.

“この不等式が成り立つように n_1, n_2 を取れば確率 1-ε で第k位のノードが出力に含まれる”

where

$\boldsymbol{P_k(n_1)}$:= 真の第k位ノードが選ばれる確率

$\boldsymbol{z_{1-\varepsilon}}$:= 平均0分散1の正規分布の(1-ε)-quantile

γ := Scale free 性を仮定。べき分布の指数**(の逆数)**

$\boldsymbol{F_k}$:= InDeg(k) $\propto k^{-\gamma}$

例: n_2=300, k=100, ε=10%, Twitter に対して計算すると $n_1 \geqq 1300$

$P_k(n_1, n_2)$

$= P[S_k \geqq S^{\dagger}_{n2}]$

$\cong P[S_k \geqq S_{n2}]$

$\cong P[\text{二項分布}(n_1, F_k/N) \geqq \text{二項分布}(n_1, F_{n2}/N)]$

$\cong P[\text{正規分布}(n_1F_k/N,\ n_1F_k/N(1-F_k/N)) \geqq \text{正規分布}(\ldots)]$

$\cong P[\text{正規分布}(n_1F_k/N,\ n_1F_k/N) \geqq \text{正規分布}(\ldots)]$

$= P[\text{正規分布}(n_1(F_k - F_{n2})/N,\ n_1(F_k + F_{n2})/N) \geqq 0]$

$= P[\text{正規分布}(\sqrt{(n_1/N)} \cdot (F_k - F_{n2})/\sqrt{(F_k + F_{n2})},\ 1) \geqq 0]$

P_k := 第k位のノードが出力に入る確率

S_k := 第k位のノードの得票数。二項分布になる

F_k := 第k位のノード入次数

正規分布で近似

正規分布の差

分散を1に

Proposition 2. *For large enough n_1, the inequality*

$$Z_k(n_1) := \sqrt{\frac{n_1}{N}} \frac{F_k - F_{n_2}}{\sqrt{F_k + F_{n_2}}} \geqslant z_{1-\varepsilon}, \qquad (14)$$

where $z_{1-\varepsilon}$ is the $(1-\varepsilon)$-quantile of a standard normal distribution, guarantees that the mean fraction of top-k nodes in W identified by Algorithm 1 is at least $1-\varepsilon$.

n_1 と n_2 の最適なバランスは？

Theorem 1. *Assume that* $k = o(n)$ *as* $n \to \infty$, *then the maximizer of the probability* $P_k(n - n_2)$ *is*

$$n_2 = (3\gamma k^\gamma n)^{\frac{1}{\gamma+1}} (1 + o(1)),$$

with γ *as in* (1).

証明: さきほど導出した P_k の近似式を元に頑張って計算。

どのくらい増やせば十分な精度？

Theorem 2. *Let the in-degrees of the entities in W be independent realizations of a regularly varying distribution G with exponent $1/\gamma$ as defined in (1), and $F_1 \geqslant F_2 \geqslant \cdots \geqslant F_M$ be their order statistics. Then for any fixed $\varepsilon, \delta > 0$, Algorithm 1 finds the fraction $1 - \varepsilon$ of top-k nodes with probability $1 - \delta$ in*

$$n = O(N/a(M))$$

API requests, as $M, N \to \infty$, where $a(M) = l(M)M^{\gamma}$ and $l(\cdot)$ is some slowly varying function.

In the case $M = N$, as in our experiments on Twitter, Theorem 2 states that the complexity of the algorithm is roughly of the order $N^{1-\gamma}$, which is much smaller than linear

- 証明: 計算。

まとめ

- 次数の高いノードを少ないクエリ回数で発見
- 非常にシンプルなアルゴリズム
- Extreme Value Theory を用いた解析

for w *in* W **do**
  $S[w] \leftarrow 0$;
for $i \leftarrow 1$ **to** n_1 **do**
  $v \leftarrow random(N)$;
  foreach w *in* $OutNeighbors(v) \subset W$ **do**
    $S[w] \leftarrow S[w] + 1$;
$w_1, \ldots, w_{n_2} \leftarrow Top\_n_2(S)$ // $S[w_1], \ldots, S[w_{n_2}]$ are the top n_2 maximum values in S;
for $i \leftarrow 1$ **to** n_2 **do**
  $d_i \leftarrow InDegree(w_i)$;