Fast Approximation of Betweenness Centrality through Sampling

Matteo Riondato and
Evgenios M. Kornaropoulos

Brown University

WSDM – February 27th 2014

Who is important?

- Key question in graph analysis:
  - Which vertices are important?

Need formal definition

- Betweenness centrality of vertex v :

  $b(v)$ = fraction of Shortest Paths going through v

Formally...

- Graph $G = (V, E)$ $|V| = n$ $|E| = m$
- Betweenness of $v \in V$:

$$\mathsf{b}(v) = \frac{1}{n(n-1)} \sum_{p_{uw} \in \mathbb{S}_G} \frac{\mathbb{1}_{\mathcal{T}_v}(p_{uw})}{\sigma_{uw}}$$

- $\mathbb{S}_G$: all SPS in G
- $\mathcal{S}_{uv}$: SPs from u to v ($\sigma_{uv} = |\mathcal{S}_{uv}|$)
- $\mathcal{T}_v = \{p \in \mathbb{S}_G \ : \ v \in \mathsf{Int}(p)\}$

Computing betweenness

- Exact algorithm [Brandes01]: $O(nm + n^2 \log n)$

too much for large networks

- Idea: use sampling!

- Goal: compute (ε, δ)-approximation
  - set of estimations $\tilde{\mathsf{b}}(v)$ for all $v \in V$ such that

$$\boxed{\Pr\left(\forall v \in V, |\tilde{\mathsf{b}}(v) - \mathsf{b}(v)| \leq \varepsilon\right) \geq 1 - \delta}$$

  - [BrandesPich07]: use $\frac{1}{\varepsilon^2}\left(\log_2 n + \ln \frac{1}{\delta}\right)$ samples

Computing betweenness

- Exact algorithm [Brandes01]: $O(nm + n^2 \log n)$

too much for large networks

- Idea: use sampling!

- Goal: compute (ε, δ)-approximation
  - set of estimations $\tilde{\mathrm{b}}(v)$ for all $v \in V$ such that

$\Pr\Big(\forall v \in V, |\tilde{\mathrm{b}}(v)$ Still too much! $\geq 1 - \delta$

- [BrandesPich07]: use $\frac{1}{\varepsilon^2}\left(\log_2 n + \ln \frac{1}{\delta}\right)$ samples

Our algorithm

- Samples shortest paths
- No. samples independent from n (VC-dimension)

$\tilde{\mathrm{b}}(v) \leftarrow 0, \forall v \in V$

For r times do

  Sample random pair (u, v)

  Compute $\mathcal{S}_{uv}$ (all SPs from u to v)

  Select a SP p uniformly at random from $\mathcal{S}_{uv}$

  For each $w \in \mathsf{Int}(p)$

    $\tilde{\mathrm{b}}(w) \leftarrow \tilde{\mathrm{b}}(w) + 1/r$

Relatively easy, see paper

Computing the sample size

- We use VC-dimension [VapnikChervonenkis71]
  - notion from Statistical Learning Theory
  - given domain D, measures "richness" of $\mathcal{R} \subseteq 2^D$
  - combinatorial property of $\mathcal{R}$
  - useful to approximate probability masses of subsets in $\mathcal{R}$ using samples from D

Sampling theorem

- Assume we know that $\mathrm{VC}(\mathcal{R}) \leq d$
- Let π be a probability distribution on D
- Given $\varepsilon, \delta \in [0,1]$, let S be a collection of samples from π
- If

$$\boxed{|S| \geq \frac{1}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right)}$$

then

Empirical Average

$$\boxed{\Pr\left(\forall A \in \mathcal{R}, \left|\pi(A) - \frac{1}{|S|}\sum_{s \in S} \mathbb{1}_A(s)\right| \leq \varepsilon\right) \geq 1 - \delta}$$

In our case...

- $D = \mathbb{S}_G$ (all shortest paths in G)
- $\mathcal{R}_G = \{\mathcal{T}_v, v \in V\}$ ($\mathcal{T}_v$ = all SPs going through v)
- Probability distribution on $\mathbb{S}_G$:

$$\boxed{\pi_G(p_{uv}) = \frac{1}{n(n-1)} \frac{1}{\sigma_{uv}}}$$

- $\pi_G(\mathcal{T}_v) = \mathsf{b}(v)$
- Our algorithm
  - samples SPs according to π_G
  - compute empirical average $\tilde{\mathsf{b}}(v), \forall v \in V$

Sample size

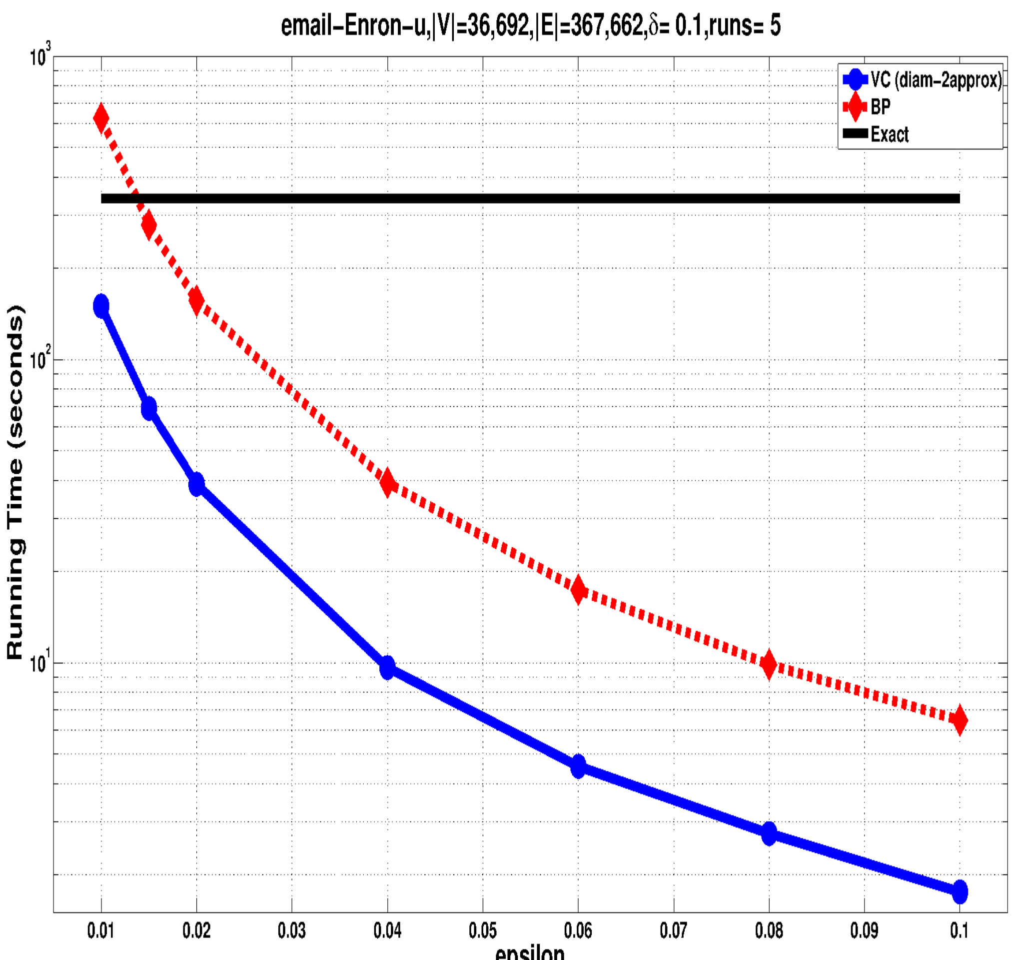
- Define vertex diameter: $\mathsf{vd}(G) = \max\{|p| \,:\, p \in \mathbb{S}_G\}$
  - can be estimated efficiently
- Theorem: $\mathsf{VC}(\mathcal{R}_G) \le \lfloor \log_2(\mathsf{vd}(G) - 2) \rfloor + 1$
- Sample size for (ε, δ)-approximation:

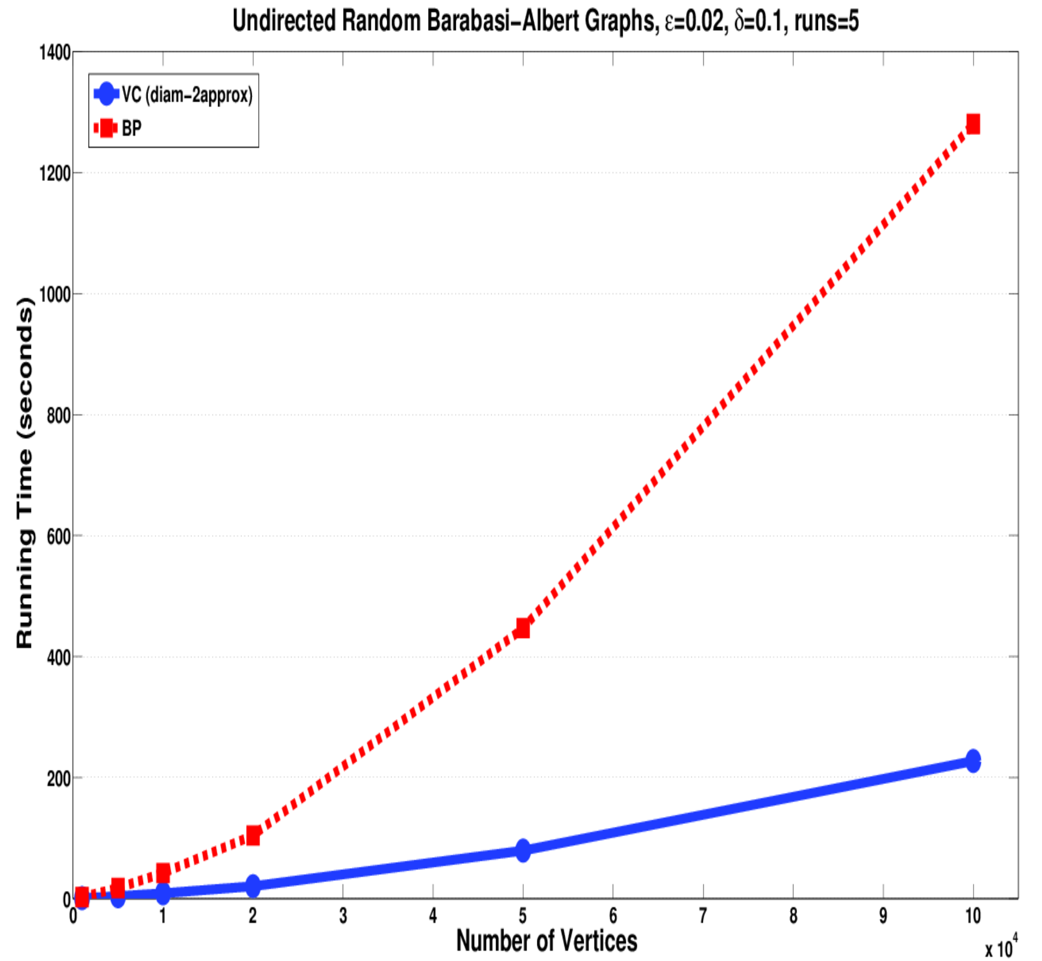
$$\frac{1}{\varepsilon^2}\left(\lfloor \log_2(\mathsf{vd}(G) - 2) \rfloor + 1 + \ln\frac{1}{\delta}\right)$$

- Independent from $|V|$

Speedup Results

- ~8x faster than [BrandesPich07]
- Scales better as n grows





I'm graduating soon!

- Looking for exciting postdoc opportunities

- Matteo Riondato
  - @riondabsd
  - http://cs.brown.edu/~matteo
  - matteo@cs.brown.edu

Bibliography

- [Brandes01] *A faster algorithm for betweenness centrality*, J. Math. Sociol., 2001

- [BrandesPich07] *Centrality estimation in large networks*, Intl. J. Bifurc. Chaos, 2007

- [EppsteinWang01] Fast approximation of centrality, J. Graph Alg. and App., 2001

- [VapnikChervonenkis71] *On the uniform convergence of relative frequencies of events to their probabilities*, Th. Prob. and its Appl., 1971