Fast Approximation of Betweenness Centrality through Sampling

Matteo Riondato and Evgenios M. Kornaropoulos ({matteo,evgenios}@cs.brown.edu)
Department of Computer Science – Brown University

Accepted as full paper at ACM WSDM'14. Full version: http://cs.brown.edu/~matteo/papers/RiondatoKornarop-BetweennessSampling.pdf . This work is supported in part by the National Science Foundation under award IIS-1247581.

I. Background

Graph $G=(V,E)$, $|V|=n$, $|E|=m$

Path p from u to v : $p=(u,\underbrace{a_1,a_2,\dots,a_\ell}_{\text{Int}(p)\ \text{Internal nodes}},v)$

For $(u,v)\in V\times V$ let $\mathcal{S}_{uv}$ = all shortest paths from u to v

Let $\mathbb{S}_G=\bigcup_{(u,v)\in V\times V}\mathcal{S}_{uv}$ = all shortest paths in G

Betweenness of $w\in V$:

$$\mathsf{b}(w)=\frac{1}{n(n-1)}\sum_{p_{uv}\in\mathbb{S}_G}\frac{\mathbb{1}_{\text{Int}(p)}(w)}{|\mathcal{S}_{uv}|}$$

It is the fraction of shortest paths that w is internal to

Measures centrality (i.e., importance) of nodes

Exact algorithms exist for computing betweenness of all vertices [Brandes01] . Too **slow** for huge graphs:

$O(n^2\log n)$ unweighted graphs, $O(nm+n^2\log n)$ weighted graphs

II. Goals and results

Key observation: In graph mining, fast and approximate algorithms are often preferred over slow but exact

Our goal: **Speed up** by computing **fast high-quality approximation** of the betweenness of all vertices using **random sampling** (sample: collection of shortest paths)

Constraints / Desiderata:

- number of samples must not depend on $|V|$
- probabilistic guarantees on quality of approximation

Results: We developed an algorithm which computes estimations $\tilde{\mathsf{b}}(v)$ for the betweenness $\mathsf{b}(v)$ of all nodes. We have that, for fixed $\varepsilon,\delta\in(0,1)$

$$\Pr(v\in V \text{ s.t.} |\tilde{\mathsf{b}}(v)-\mathsf{b}(v)|>\varepsilon)<\delta$$

i.e.,all estimations are very accurate with high probability

III. Our algorithm

1) $\forall u\in V$, let $\tilde{\mathsf{b}}(v)=0$
2) For $i=1,\dots,r$ ← Sample Size
3) Sample random pair (u,v) of different nodes
4) Sample random shortest path p_i from $\mathcal{S}_{uv}$
5) For $w\in\text{Int}(p_i)$, $\tilde{\mathsf{b}}(w)=\tilde{\mathsf{b}}(w)+1/r$
6) Return $\{\tilde{\mathsf{b}}(u),u\in V\}$

Path p_{uv} is sampled with probability $\pi_{p_{uv}}=\dfrac{1}{n(n-1)|\mathcal{S}_{uv}|}$

Sampling a shortest path is **easy**: perform Dijkstra from u to v, then backtrack choosing predecessor z at random with probability proportional to the number of shortest paths in $\mathcal{S}_{uv}$ that z is internal to

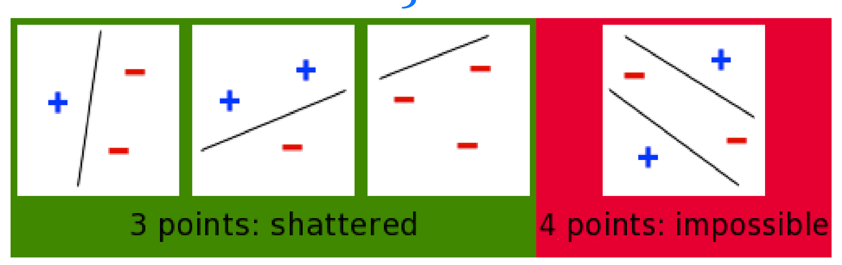
Choice of sample size r is **crucial for correctness**

We use VC-dimension, a core notion from Statistical Learning Theory to compute it

IV. VC-Dimension

Definition: Given a **set of points** P and a **family** $F\subseteq 2^P$ the **VC-Dimension** of (P,F) is the **cardinality of the largest** $A\subseteq P$ such that $\{R\cap A\ :\ R\in F\}=2^A$

Example
$P=\mathbb{R}^2$, F = oriented halfspaces
VC-Dimension of (P,F) = 3



Theorem (Bound to sample size)

Let (P,F) have VC-dimension $\leq d$, and let π be a probability distribution on P. Given $\varepsilon,\delta\in(0,1)$ let S be a collection of points from P sampled according to π with size

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

If d does not depend on |P|, then **|S| is also independent from |P|!**

Then

$$\Pr\left(f\in F \text{ s.t. } \left|\mathbb{E}_\pi[f]-\frac{\sum_{c\in S}\mathbb{1}_f(c)}{|S|}\right|>\varepsilon\right)<\delta$$

i.e., the empirical averages for all $f\in F$ are close to their expectations

A sample S for which the above event is verified is called an **ε-sample for** (P,F)

V. The vertex-diameter of the graph

For $u\in V$, let $T_u=\{p\in\mathbb{S}_G\ :\ u\in\text{Int}(p)\}$

In our case: $P=\mathbb{S}_G$, $F=\{T_u,u\in V\}$

If $S=\{p_1,\dots,p_r\}$ (sampled paths) is ε-sample for $(\mathbb{S}_G,F)$ then

$$\left|\underbrace{\mathbb{E}_\pi[T_u]}_{\mathsf{b}(u)}-\underbrace{\frac{\sum_{p\in S}\mathbb{1}_{T_u}(p)}{|S|}}_{\tilde{\mathsf{b}}(u)}\right|\leq\varepsilon\quad\forall u\in V$$

since $\mathbb{1}_{T_u}(p)=\mathbb{1}_{\text{Int}(p)}(u)$

We only need a bound to VC-dimension of $(\mathbb{S}_G,F)$!

Definition:The vertex-diameter of G is $\mathsf{vd}_G=\max_{p\in\mathbb{S}_G}|\text{Int}(p)|$

An approximation of vd_G can be computed efficiently.

vd_G is small & shrinks in social networks as size grows!

Thm: VC-dimension of $(\mathbb{S}_G,F)$ is at most $\lfloor\log_2\mathsf{vd}_G\rfloor+1$

Theorem: We need to sample r shortest paths with

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

The sample size is independent from $|V|$

VI. Results of experimental evaluation

We used graphs from SNAP (http://snap.stanford.edu)

Very accurate estimation: $|\tilde{\mathsf{b}}(u)-\mathsf{b}|\ll\varepsilon,\forall u\in V$

Very fast execution, even for small ε

Better scalability than existing solution [BrandesP08]