Large Very Dense Subgraphs in a Stream of Edges

Claire Mathieu
IRIF-CNRS

Michel de Rougemont
University Paris II
IRIF-CNRS
Plan

1. Stream of graph edges
   • Hard problems: Maxclique, \((\gamma, \delta)\)-cluster
   • special case: social graphs

2. Context: giant components of random graphs
   • Erdos-Renyi model
   • Power law degree distribution and configuration model: Molloy-Reed
   • Our algorithm: keep \(k\) uniform sampled edges, observe the giant components

3. Main result: 1-way stochastic approximation: Detection of a \((\gamma, \delta)\)-cluster
   • If \(G\) has such a cluster and \(k = \Theta(\sqrt{n \log n})\), the algorithm accepts with h.p.
   • For a random input on \(\mu\), if \(G\) does not have such a cluster, the algorithm rejects with h.p.

4. Other results: Lower bound, Reconstruction, Extensions to dynamic graphs

Conclusion: Finding a \((\gamma, \delta)\)-cluster is not so hard on social graphs
1. Stream of graph edges \( e_1, e_2, \ldots, e_m, \ldots \)

\[ S \text{ is a } (\gamma, \delta) \text{- cluster if:} \]
- \( |E(S)| \geq \gamma |S| |S - 1|/2 \)
- \( |S| \geq \delta \sqrt{n} \)

MaxClique has value \( \sqrt{n} \) iff there is a \((1,1)\)-cluster

Hard problem [Hastad 1999]:
No poly-time \( n^{0.99} \) approximation of MaxClique unless P=NP.

**Goal of the paper:** existence of a \((\gamma, \delta)\)-cluster
is not so hard on social graphs
A social graph: Twitter Graph

Twitter Graph $G$

@JoeBiden: With @KamalaHarris. Make sure to vote for #Election2020.

Nodes=$\{\@JoeBiden, \@KamalaHarris, \#Election2020\}$

Edges=$\{ (\@JoeBiden, \@KamalaHarris), (\@JoeBiden, \#Election2020)\}$

**Observation:** $G$ has a heavy-tailed degree distribution

**Hypothesis:** $G$ follows a power law degree distribution

Degree sequence $D=(c.n, c.n/4, c.n/9,\ldots)$

\[
\begin{align*}
c & \simeq 0.6 \\
m &= cn \log n/4 \\
degree_{\text{max}} &= \sqrt{c.n}
\end{align*}
\]

\[
\text{Prob}[\text{degree}(u) = i] = \frac{c}{i^2}
\]
Twitter graphs have $(\gamma, \delta)$-clusters

Twitter with $m=10^4$ edges. We see clusters.  

$\gamma$,$\delta$-clusters $k=500$ uniform random edges

Observation: clusters in $G$ seem to correspond to large connected components in $R$
Reservoir sampling (k)

Q: How do we get $k$ uniform random edges in a graph given as a stream of edges?

A: Reservoir sampling [Vitter 80’s]: first store $e_1, e_2, ..., e_k$ in $R$
   for all $i > k$, store $e_i$ in $R$ with probability: $k/i$
   Replace a random $e_j$ in $R$ by $e_i$

Detection algorithm to answer the question “does $G$ have a $(\gamma, \delta)$-cluster?”:

- Reservoir sampling $R$ of size $k = \Theta(\sqrt{n \log n})$
- Observe the giant components of $R$

Output YES if $R$ has a large enough connected component, NO otherwise.

Next task: analyze our algorithm
2. Random graphs & Giant Components

1. ER: Erdös-Renyi $G(n,p)$
   sampling the complete graph $p=k/m$ produces a sample with $k$ edges on average
   extension: sampling on $\gamma$-cluster $p>1/\gamma.n \Rightarrow$ giant component

1. CM: Configuration Model [Bollobás 80] $\mu$ creates random graph with given degree distribution,

Degree distributions: [Molloy-Reed 2008] give sufficient conditions $\Rightarrow$ giant comp.

3. Our model: CM | ER
   With CM, generate a graph with a power law degree distribution $D$
   Then take uniform samples ($k$ edges)
3. Main result

Detection Algorithm $A(\gamma, \delta)$
- Reservoir Sampling $k = \frac{c.\sqrt{n}.\log n}{4.\gamma.\delta}$
- Let $C$ be the largest connected component
  If $|C| \geq \lambda = \Theta(n^{1/8}.\log^2 n)$ Accept, else Reject

1-way stochastic Approximation ($\mu$)

Lemma 1. If $G$ has a ($\gamma, \delta$)-cluster, then $A$ accepts with h.p.

Theorem 1. If $G$ is a random graph from $\mu$ with no ($\gamma, \delta$)-cluster, $A$ rejects with h.p.
If G has a \((\gamma, \delta)\)-cluster

Lemma 1. Let G have \(m = cn\log n/4\) edges. If G has a \((\gamma, \delta)\)-cluster, then there is a giant component in the Reservoir with h.p.

Proof: Reservoir(k) : Erdös-Renyi G(n,p) \(p = k/m\)

\[\exists S \text{ s.t. } |S| \geq \delta.\sqrt{n}\]

\[
\frac{k}{m} = \frac{c.\sqrt{n}.\log n}{4\gamma.\delta} \cdot \frac{4}{c.n.\log n} = \frac{1}{\gamma.\delta.\sqrt{n}} \geq \frac{1}{\gamma.|S|}
\]

Recall: \(p > 1/\gamma.n \Rightarrow \text{giant component}\)

Conclusion: there is a giant component in R, and so, A accepts w.h.p.
If G is a random graph from $\mu$:

Lemma 2. W.h.p. G has no $\gamma$-cluster of size $\Omega(\sqrt{n})$. (Proof omitted)

Proof of Theorem 1: If G is a random graph from $\mu$ with no $(\gamma, \delta)$-cluster, A rejects with h.p.

Molloy-Reed (2008) give sufficient conditions on a degree distribution $D$ for the configuration model to have no giant component w.h.p.: if

- D is “well-behaved”
- $Q(D) = E(D^2) - 2E(D) < 0$
- Conditions on maximum and average degree then $|\text{largest connected component}| < b \cdot n^{1/4}$
Analysis of degree distribution $D_R$ in $R$

**Difficulty:** $D_R$ is probabilistic

First, analyze $E(D_R)$ to prove the Molloy Reed conditions

- $E(D_R)$ is well behaved with h.p (uniform convergence, ....)
- Maximum degree and Average degree conditions
- $Q(E(D_R)) < 0$

Second, modify the probability space
Configuration: first and last

Configuration Model
D=(2,4)

Reservoir
k=2
Configuration:last: sample first, then match

Analysis with h.p. of the Molloy Reed conditions

- $D_R$ is well behaved with h.p (uniform convergence,....)
- Maximum degree and Average degree conditions
- $Q(D_R) < 0$

Goal: produce a deterministic degree sequence
Sketch of the proof of theorem 1

If G is a random graph from μ with no (γ,δ)-cluster, A rejects with h.p.

Consider a degree sequence coupling degree i and n. Apply Molloy-Reed, deduce bound on size of max connected component C.

\[ P_{\text{configuration last}}[|C| \leq k^{1/4}] = \\
\lim_{n \to \infty} P_{\mu,\Omega}[|C| \leq k^{1/4}] = 1 \]

Thus R has no giant component with h.p.
Recall Lemma 2: G has no (γ, δ)-cluster w.h.p.
**Conclusion:** Detection algorithm is correct with h.p.
4. Other result (1) : Space lower bound

Multiparty Disjointness Problem (n,q): q parties, 1-way communication, DISJ(n,q)

Bahmani et al. 2012: BKV-reduction

\[ \text{DISJ}(n, \sqrt{n}) \prec \exists(\gamma, \delta) - \text{cluster} \]

\[
\begin{array}{c|c|c}
1 & 0 & 0 \\
0 & 0 & 0 \\
0 & 1 & 1 \\
\end{array}
\]

\[ n = 4, q = 3 \]

Theorem 2: Any algorithm to decide whether there exists a cluster requires \( \Omega(\sqrt{n}) \) space.
Other result (2): Reconstruction algorithm

Assume that $G$ has a clique ($\gamma=1$) of size $\Omega(\sqrt{n})$.

Q: Can we reconstruct the Clique from the Reservoir?
A: Output 2-core(largest connected component(Reservoir))
**Other result (3): Dynamic graphs**

Sliding windows (old edges disappear)
Reservoirs for each window

**Dynamic Algorithm:** keep the large connected components of the Reservoirs for each window.

**Goal:** measure the changes in the giant components.
Conclusion

Problem: Existence of a \((\gamma,\delta)\)-cluster, Maxclique

Not so hard for social graphs.

Main result: Streaming algorithm with space \(k = \Theta(\sqrt{n \log n})\)

Main notion: 1-way stochastic approximation(\(\mu\)):
If \(G\) has a \((\gamma,\delta)\)-cluster, then \(A\) accepts with h.p.
If \(G\) is a random graph from \(\mu\) with no \((\gamma,\delta)\)-cluster, \(A\) rejects with h.p.