Coreset and sampling approaches for the analysis of very large data sets – Part II

Christian Sohler
Very Large Networks

Social Networks

- Reflect social structures in detail
- **Example question:** Can we distinguish democratic countries from totalitarian ones by looking at their Facebook structure?

Data size

- GigaByte upto TeraByte (only the graph)
- Data exchange (movies, pictures, etc.) in the Peta-Byte range

Source: [1]
Very Large Networks

Problem
- Such networks are typically too large to be compared
- No efficient algorithm for graph isomorphism
- In order to apply learning algorithms, we need features that describe different aspects of the network structure

Source: [1]
Sublinear Algorithms

Observations
- Classical algorithms are too slow to handle very large networks
- In some learning applications we want to be able to handle many features of many large networks

Property Testing
- Study (structural) properties of very large networks via random sampling
- A form of approximation
- Central question: What can we provably learn from the local structure of a graph about its global structure?
Graph Properties

Graph
- Graph $G=(V,E)$, $V=\{1,\ldots,n\}$
- Bounded max. degree $D$

Definition (graph property)
- A **graph property** is a set of graphs that is closed under isomorphism.

Definition (ε-far)
- A graph $G=(V,E)$ is **ε-far from a property $P$**, if one has to modify more than $\varepsilon Dn$ edges to obtain a degree bounded graph with property $P$.
- If a graph is not $\varepsilon$-far from $P$, it is called **ε-close**.
Property Testing [Rubinfeld, Sudan, SICOMP 96]

Property Tester for P [Goldreich, Ron, Algorithmica 02]

- Oracle access to graph $G=(V,E)$:
  - Query$(i,j)$ returns $i$-th edge incident to vertex $j$ or a symbol that this edge does not exist
- Accepts with prob. at least $2/3$, if $G$ has property $P$
- Rejects with prob. at least $2/3$, if $G$ is $\varepsilon$-far from $P$

Quality measures

- Query complexity: maximal number of oracle queries
- Running time
A Simple Example: Connectivity

Connectivity
- Every vertex is connected (has a path) to every vertices
- $\varepsilon$-far: There are at least $\varepsilon Dn/2$ connected components

Connectivity tester ($\varepsilon$) [Goldreich, Ron, Algorithmica 02]
(1) Sample set $S$ with $s=O(1/\varepsilon)$ vertices uniformly at random from $V$
(2) For every vertex from $S$:
(3) Perform a BFS until
   (a) $4/(\varepsilon D)$ vertices have been discovered or
   (b) all vertices of a small connected component have been discovered
      if (b) then reject
(4) accept
## Two Main Sampling Approaches

**Frequent subgraph analysis**

1. Sample set $S$ of vertices uniformly at random

2. For each vertex in $S$ determine subgraph induced by vertices within distance at most $k$

3. Decide based on the observed subgraph

* Requires bounded max. degree

**Random walks**

1. Sample set $S$ of vertices uniformly at random

2. From each vertex in $S$ start a $t$-step random walk

3. Decide based on the observed subgraph
General Question

Definition
- A graph property $P$ is called testable, if there is a $q=q(\varepsilon, D)$, such that for every $n>0$ and every $\varepsilon, 0<\varepsilon \leq 1$, a property tester $A_{\varepsilon,D,n}$ with query complexity $q$ exists.

General question
- Which properties are testable with constant query complexity for constant $\varepsilon$?
Frequent Subgraph Analysis

Frequent Subgraph Analysis
1. Draw sample set \( S \subseteq V, |S| = s(\epsilon, D) \), uniformly at random
2. Let \( k = k(\epsilon, D) \)
3. Accept, if all \( k \)-balls \( H(k, v) \) have the studied property
4. Reject otherwise

Definition (k-ball)
A \( k \)-ball \( H(k, v) \) around a root vertex \( v \) in a graph \( G \) is the subgraph induced by all vertices of distance at most \( k \) from \( v \).
Simplified General Question

**Simplified question** [Czumaj, Shapira, Sohler, SICOMP 09]

- Which properties are testable with constant query complexity, if the input graph is planar?
- Planarity is an example for a larger class of graphs
Simplified General Question

Simplified question [Czumaj, Shapira, Sohler, SICOMP 09]

- Which properties are testable with constant query complexity, if the input graph is planar?
- Planarity is an example for a larger class of graphs

How does planarity help?

- Every degree bounded planar graph can be partitioned into connected components of size $O(1/\varepsilon^2)$ by removing at most $\varepsilon Dn/2$ edges
- If a graph is $\varepsilon$-far from $P$, then it is $\varepsilon/2$-far after the removal of these edges
Simplified General Question

**Theorem** [Czumaj, Shapira, Sohler, SICOMP 09]
- In the class of planar graphs every graph property that is closed under vertex removal is (non-uniformly) testable.

**Proof idea (simplified)**
- Use frequent subgraph analysis
- \( G \) has \( P \): The tester accepts by closedness under vertex removal
- \( G \) is \( \varepsilon \)-far from \( P \): After removal of \( \varepsilon Dn/2 \) edges, \( G \) has small connected components and is \( \varepsilon/2 \)-far from \( P \)
- Hence, there are many components that do not have property \( P \)
- With constant probability a random \( k \)-ball contains such a component
- Because of closedness the \( k \)-ball does not have \( P \) and the tester rejects
What FrequentSubgraphAnalysis1 Cannot Do

PlanarityTesterFirstTry($\epsilon,n$)

1. Draw $s = s(\epsilon,D)$ k-balls uniformly at random for a $k = k(\epsilon,D)$
2. Accept, if only planar k-balls are drawn and reject, otherwise
What FrequentSubgraphAnalysis cannot do

PlanarityTesterFirstTry(ε,n)
(1) Draw s=s(ε,D) k-balls uniformly at random for a k=k(ε,D)
(2) Accept, if only planar k-balls are drawn and reject, otherwise

Counter example:
- There are classes of graphs, such that every cycle has length $\Omega(\log n)$ and that are $\varepsilon$-far from planar
Extended Frequent Subgraph Analysis

Frequent Subgraph Analysis II
1. Draw sample set \( S \subseteq V, |S|=s(\epsilon, D) \), uniformly at random
2. Let \( k=k(\epsilon, D) \)
3. Accept, based on the frequency of the observed \( k \)-balls and internal randomness

Observations
- For constant \( k \) and \( D \) there is only a constant number of non-isomorphic \( k \)-balls
- The distribution vector \( \text{freq}(G,k) \) of the \( k \)-balls in a graph \( G \) describes the relative frequency of \( k \)-balls in \( G \)
Back to the General Question

**Theorem** [Benjamini, Schramm, Shapira, Advances in Mathematics 10]
- Every graph property that is closed under removal of vertices, removal of edges and contraction of edges is (non-uniformly) testable.
Back to the General Question

**Theorem** [Benjamini, Schramm, Shapira, Advances in Mathematics 10]
- Every graph property that is closed under removal of vertices, removal of edges and contraction of edges is (non-uniformly) testable.

**Getting some intuition**
- We will first try to distinguish expander graphs from planar graphs

**Two definitions**
- Define the *conductance* of a set of vertices $U$ to be $|E(U, V-U)| / |U|$.
- A graph is called an *expander graph*, if every set $U$ of vertices with $|U| \leq |V|/2$ has conductance $\Omega(1)$
Some intuition

Comparing trees and high girth expander graphs
Some intuition

Comparing trees and high girth expander graphs
Some intuition

Planar graphs
- Planar graphs can be decomposed into connected components of size at most k by removing at most $Dn/(2\sqrt{k})$ edges
- On average such a connected component is incident to $O(\sqrt{k})$ removed edges and so it has conductance $1/\sqrt{k}$

Expander graphs
- Conductance is $\Omega(1)$

Conclusion
- We can distinguish expander graphs from planar graphs by considering the local conductance
Back to the General Question

**Theorem** [Benjamini, Schramm, Shapira, Advances in Mathematics 10]
- Every graph property that is closed under removal of vertices, removal of edges and contraction of edges is (non-uniformly) testable.

**Proof idea**
- Use an approximation of the frequency vector of k-balls to test, whether the graph is **hyperfinite**, i.e. can be decomposed into small components (note: Every graph in the class above can be decomposed!)
- Continue with ideas of the previous algorithm
Back to the General Question

**Theorem** [Hassidim, Kelner, Nguyen, Onak, FOCS 09]
- Every (non-degenerate) hereditary property can be tested in hyperfinite graphs.

**New contributions**
- Explicit algorithm to compute partition into small components
- Improved query complexity
- Simplified proof
Back to the General Question

GlobalPartitioning(k, δ) [Hassidim, Kelner, Nguyen, Onak, 09]

- \( \pi = (\pi_1, \ldots, \pi_n) \) = random permutation of the vertices
- \( P = \emptyset \)
- while \( G \) is not empty do
  - Let \( v \) be the first vertex in \( G \) according to \( \pi \)
  - if there exists a \((k, \delta)\)-isolated neighborhood of \( v \) in \( G \) then
    - \( S = \) this neighborhood
  - else \( S = \{v\} \)
  - \( P = P \cup \{S\} \)
  - remove vertices in \( S \) from the graph

\((k, \delta)\)-isolated neighborhood of \( v \): Connected set \( S \) of size at most \( k \) with \( v \in S \) and at most \( \delta |S| \) edges between \( S \) and \( V \)
Back to the General Question

**Definition [Hassidim, Kelner, Nguyen, Onak, 09]**
We say that $O$ is a (randomized) $(\epsilon,k)$-partitioning oracle, if given query access to a planar graph $G=(V,E)$, it provides query access to a partition $P$ of $V$. For a query about $v \in V$, $O$ returns $P[v]$. The partition has the following properties:

- $P$ is a function of the graph and the random bits
- For every $v \in V$, $|P[v]| \leq k$ and $P[v]$ induces a connected graph in $G$
- $|\{(v,w) \in E : P[v] \neq P[w]\}| \leq \epsilon \cdot |V|$ with prob. $9/10$
Back to the general question

Lemma [Variant of Lemma by Hassidim, Kelner, Nguyen, Onak, 09]
Let $G$ be a planar graph with degree bounded by $D \geq 2$. Let $R = R(\varepsilon, D)$ be any function and let $S$ be a set of $|S| = R$ vertices chosen uniformly at random. Then there is a $k = k(\varepsilon)$ such that there is an $(\varepsilon D, k)$-partitioning oracle that inspects a $D = D_R(\varepsilon, D)$-ball of every vertex in $S$ and with probability $9/10$ returns the partition class (and component) of every vertex in $S$. 
Back to the General Question

**Lemma** [Newman, S. 2013]

- Let $G$ be any graph with maximum degree $D$. We can estimate the frequency vector $\text{freq}(G,k)$ up to $l_1$-error $\varepsilon$ by sampling $Q = f(\varepsilon, k)$ vertices uniformly at random, explore their $k$-discs, and return the relative frequencies of the sampled discs. We call this algorithm $\text{EstimateFrequencies}$. 
Back to the General Question

**Theorem** [Newman, Sohler, STOC 11]

- Let $G$ and $H$ be two hyperfinite (planar) graphs with $n$ vertices and max.
  degree $D$. Then for every $\varepsilon, 0 < \varepsilon \leq 1$, there is $\lambda = \lambda(\varepsilon, D)$ and $k = (\varepsilon, D)$, such that:

  If $||\text{freq}(G, k) - \text{freq}(H, k)||^1 \leq \lambda$ then $G$ $\varepsilon$-close to $H$. 
Back to the General Question

Outline (first idea)
- Let $G$ and $H$ be two graphs on $n$ vertices with $\text{freq}(G,k) = \text{freq}(H,k)$ for sufficiently large $k$
- (1) Use $(\varepsilon,k')$-partitioning oracle on $G$ and $H$ resulting in graphs $G^*$ and $H^*$
- (2) Show that $\text{freq}(G^*,k')=\text{freq}(H^*,k')$
- Wrapup: By removal of at most $\varepsilon n$ edges we obtain two graphs with the same number of isomorphic connected components

Problem
- There is no reason to expect $G^*$ and $H^*$ to have the same frequency vectors
Back to the General Question

Outline (second idea)

- Use *probabilistic method* to prove that for some choice of permutations of the vertex sets of G and H the partitioning oracle will compute graphs G* and H* with $\text{freq}(G^{*},k{'}^{\prime}) \approx \text{freq}(H^{*},k{'}^{\prime})$
Back to the General Question

Theorem [Newman, Sohler, STOC 11]

- Let $G$ and $H$ be two hyperfinite (planar) graphs with $n$ vertices and max. degree $D$. Then for every $\varepsilon, 0 < \varepsilon \leq 1$, there is $\lambda = \lambda(\varepsilon, D)$ and $k = (\varepsilon, D)$, such that:

$$\text{If } ||\text{freq}(G,k) - \text{freq}(H,k)||^{1} \leq \lambda \text{ then } G \varepsilon\text{-close to } H.$$ 

Corollary [Newman, Sohler, STOC 11]

- Every property is (non-uniformly) testable in the class of hyperfinite graphs
- Every hyperfinite graph property is (non-uniformly) testable
Challenges

Expander graphs
- Hyperfinite graphs are the “opposite“ of expander graphs
- Social networks are typically not hyperfinite (small world phenomenon)

Degree bound
- Frequent subgraph analysis requires bounded degree
- Which classes of properties can be tested in graphs with small average degree?

Directed graphs (when edges are seen from one side)
- How to analyze random walks? How to avoid „to get stuck“?
- Testable properties?
Challenges

Query complexity / running time
- The general results have a poor dependence on $\varepsilon$
- Which properties can be tested with polynomial (linear) query complexity (running time)?
- Can planarity be tested in polynomial query complexity?

One-sided vs. Two-sided error
- The general results have two-sided error
- In many applications, one wants a „counter-example“, if the graph does not have a property
- Which properties can be efficiently tested with one-sided error?
Thank you!

Image sources:
[1] TonZ; Image under Creative Commons License