Optimization in very large graphs

LÁSZLÓ LOVÁSZ
Eötvös Loránd University, Budapest

(joint results with Balázs Szegedy)
Very large graphs

- Internet
- Social networks
- Ecological systems
- VLSI
- Statistical physics
- Brain
What properties to study?

- Does it have an even number of nodes?

- How dense is it (average degree)?

- Is it connected?

- Find connected components.
• Graph is HUGE.

• Not known explicitly (not even number of nodes).
- We can sample a uniform random node a bounded number of times, and see edges between sampled nodes.

**Works in the dense case only** ($\sim cn^2$ edges)
- We can sample a uniform random node a bounded number of times, and explore its neighborhood to a bounded depth.

- Works in the sparse case: Bounded degree ($\leq d$).
Different types of algorithmic questions

- Estimate a parameter (triangle density, density of max cut, rank of the adjacency matrix,…)

- Test a property (planar, bipartite, triangle-free,…)

- Find the structure (connected components, max cut, max matching,…)

The distance of two graphs

(a) \( V(G) = V(G') \)

\[
d_{\square}(G, G') = \max_{S, T \subseteq V(G)} \frac{|e_G(S, T) - e_{G'}(S, T)|}{n^2}
\]

(b) \( |V(G)| = |V(G')| \)

\[
\delta^*_{\square}(G, G') = \min_{G \leftrightarrow G'} d_{\square}(G, G')
\]
The distance of two graphs

(c) $|V(G)| = n, \ |V(G')| = n'$

Blow up nodes:

$$\delta_{\square}(G, G') = \lim_{k \to \infty} \delta^*_{\square}(G(kn'), G'(kn))$$
The distance of two graphs

\[ d_{\square}(G, G') = \max_{S, T \subseteq V(G)} \frac{|e_G(S, T) - e_{G'}(S, T)|}{n^2} \]

\[ \delta^*_\square(G, G') = \min_{G \leftrightarrow G'} d_{\square}(G, G') \]

\[ \delta_{\square}(G, G') = \min_{G(\mathbf{k}n'), G'(\mathbf{k}n))} d_{\square}(G(\mathbf{k}n'), G'(\mathbf{k}n)) \]
The distance of two graphs

Examples: \[ \delta_{\square}(K_{n,n}, \mathcal{G}(2n, \frac{1}{2})) \approx \frac{1}{8} \]

\[ \delta_{\square}(\mathcal{G}_1(n, \frac{1}{2}), \mathcal{G}_2(n, \frac{1}{2})) = O(1) \]

Two graphs are "close" in the \( \delta \) distance \( \iff \) their subgraph distributions are "close".

Borgs-Chayes-L-Sós-Vesztergombi
Triangle density: easy

Maximum cut: nontrivial
The maximum cut problem

maximize

NP-hard, even with 6% error
Hastad
Polynomial-time computable with \( \approx 13\% \) error
Goemans-Williamson

Applications: optimization, statistical mechanics…
Density of maximum cut

cut with many edges

cut with many edges
A graph parameter $f$ can be estimated from samples if and only if

(i) $\forall \varepsilon > 0 \ \exists \delta > 0$ s.t. $V(G) = V(G')$ and $d(G, G') < \delta$ 

$$\implies |f(G) - f(G')| < \varepsilon.$$

(ii) $|f(G) - f(G - v)| \to 0 \quad (|V(G)| \to \infty)$

(iii) $\forall G: f(G(m))$ is convergent as $m \to \infty.$

Borgs, Chayes, L, Sós, Vesztergombi
“Property testing”: Arora-Karger-Karpinski

Goldreich-Goldwasser-Ron

Rubinfeld-Sudan

Fischer

Frieze-Kannan

Alon-Shapira
The key to algorithmic results in the dense case

Original Regularity Lemma  
Szemerédi 1976

“Weak” Regularity Lemma  
Frieze-Kannan 1999

“Strong” Regularity Lemma  
Alon – Fisher – Krivelevich - M. Szegedy
$G$: graph

$P=\{V_1, \ldots, V_k\}$: partition of $V(G)$

$G_P$: edge-weighted complete graph on $V(G)$, where the weight of edge $uv$ ($u \in V_i$, $v \in V_j$) is

$$p_{ij} = e_G(V_i, V_j)/|V_i||V_j|$$
“Weak" Regularity Lemma (Frieze-Kannan):

\( \forall k \geq 1, \forall \text{ graph } G \exists \text{ partition } P = \{ V_1, \ldots, V_k \} \) such that

\[
d_\square (G, G_P) \leq \frac{4}{\sqrt{\log k}}
\]
Two nodes are "similar", if they are connected.

Does not measure what we need...

They are similar, if their neighborhoods are (almost) the same.

Too strong...

See: random graph
Similarity distance of nodes

\[ d_2(s, t) := E_v \left| P_u(s, v \in N(u)) - P_w(t, v \in N(w)) \right| \]

**Fact 1:** This is a metric.

**Fact 2:** Can be computed by sampling.
Representative set of nodes

(i) $u, v \in R \Rightarrow d_2(u, v) > \varepsilon$

(ii) $u \in V(G) \Rightarrow d_2(u, R) \leq \varepsilon$

Any maximal set with (i) will do.
Representative set of nodes

(i) $u, v \in R \Rightarrow d_2(u, v) > \varepsilon$

(ii) $u \in V(G) \Rightarrow d_2(u, R) \leq \varepsilon$

Every graph contains an approximate representative set with at most $2^{2/\varepsilon^2}$ elements.
Representative set – Voronoi diagram

Voronoi diagram
= weak regularity partition
Representative set – Voronoi diagram

\[ S \subseteq V(G) : \quad \bar{d}(S) = \mathbb{E}_x d_2(x, S) \quad \text{average } \epsilon\text{-net} \]

\[ \mathcal{P} \text{ partition: } \ r(\mathcal{P}) = \delta \sqcap (G, G_{\mathcal{P}}) \quad \text{regular partition} \]

Voronoi cells of \( S \) form a partition with

\[ r(\mathcal{P}) < 8\sqrt{\bar{d}(S)} \]

\[ \forall \text{ partition } \mathcal{P} = \{V_1, \ldots, V_k\} \text{ of } [0,1] \quad \exists \ v_i \in V_i \text{ with } \]

\[ \bar{d}(\{v_1, \ldots, v_k\}) < 8r(\mathcal{P}) \]
Representative set – algorithm

- Begin with $U=\emptyset$.
- Select random nodes $v_1, v_2, \ldots$
- Add $v_i$ to $U$ iff $d_2(v_i, u) > \varepsilon$ for all $u \in U$.
- Stop if for more than $1/\varepsilon^2$ trials, $U$ did not grow.

size bounded by $2^{2/\varepsilon^2}$
In which class does node $v$ belong?

Let $U = \{u_1, \ldots, u_k\}$.

Put node $v$ in $V_i$ iff $i$ is the first index with $d_2(u_i, v) \leq \varepsilon$. 

Representative set – algorithm
Max cut – algorithm

Constructing representation of cut:

- Construct representative set $U$
- Compute $p_{ij} = \text{density between classes } V_i \text{ and } V_j$ (use sampling)
- Compute max cut $(U_1, U_2)$ in complete graph on $U$ with edge-weights $p_{ij}$
On which side of the cut does $v$ belong?

Put node $v$ of left side of cut iff

$$d_2(U_1,v) \leq d_2(U_2,v).$$

(Different algorithm implicit by Frieze-Kannan.)
Every graph contains a representative set with at most $2^{2/\varepsilon^2}$ elements.

Typically there is a much smaller one.

(i) $u, v \in R \Rightarrow d_2(u, v) > \varepsilon$

(ii) $u \in V(G) \Rightarrow d_2(u, R) \leq \varepsilon$

Looks like dimension.
Convergence and limit objects

$t(F, G)$: Probability that random map $V(F) \rightarrow V(G)$ preserves edges

$(G_1, G_2, \ldots)$ convergent: $\forall F \ t(F, G_n)$ is convergent

distribution of $k$-samples is convergent for all $k$
A graph parameter $f$ can be estimated from samples if and only if

$$(G_n) \text{ convergent } \Rightarrow f(G_n) \text{ convergent}$$

Borgs, Chayes, L, Sós, Vesztergombi
\[ \mathcal{W}_0 = \{ W: [0,1]^2 \rightarrow [0,1], \text{ symmetric, measurable} \} \]

(graphons)

\[ t(F, W) = \int_{[0,1]^{\gamma(F)}} \prod_{ij \in E(F)} W(x_i, x_j) \, dx \]

\[ G_n \rightarrow W : \forall F: t(F, G_n) \rightarrow t(F, W) \]
For every convergent graph sequence \((G_n)\) there is a \(W \in \mathcal{W}_0\) such that \(G_n \rightarrow W\).

Conversely, \(\forall W \ \exists (G_n)\) such that \(G_n \rightarrow W\).

\(W\) is essentially unique

(up to measure-preserving transform).

L – B. Szegedy

Borgs – Chayes - L
The distance $\delta$ between graphons, the distance $d_2$ between points, representative sets, regularity partitions,.... can be defined for graphons $(W_0, \delta)$ is a compact metric space.

The completion of $([0,1],d_2)$ is a compact metric space for every graphon.
The distance $\delta$ between graphons, the distance $d_2$ between points, representative sets, regularity partitions, .... can be defined for graphons

$(\mathcal{W}_0, \delta)$ is a compact metric space.

The completion of $([0,1], d_2)$ is a compact metric space.

Equivalent to all versions of the Regularity lemma.
If \([0,1],d_2\) has finite dimension for some graphon \(W\), then \(\forall \varepsilon\) it has a representative set/weak regularity partition with \((1/\varepsilon)^{\text{const}}\) elements.

If \(G\) is a graph that does not contain \(F\) as a bipartite-induced subgraph (\(F\) bipartite), then \(\forall \varepsilon\) it has a representative set/weak regularity partition with \((1/\varepsilon)^{10|F|}\) elements.