Testing dense graphs

- bipartiteness

Adjacency Matrix model:

G represented by matrix A
s.t. can query A in one step

$$A = \boxed{\quad A_{ij} \quad}$$

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{o.w.} \end{cases}$$

Distance from property P:

<u>def</u> G is ε-far from P if must change $> \varepsilon n^2$ entries in A to turn G into a member of P

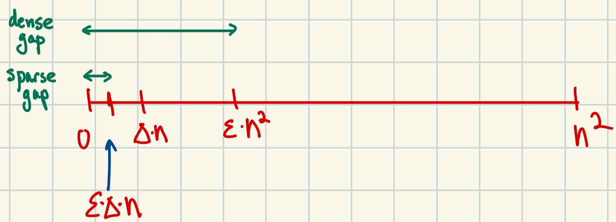
testing "sparse" properties e.g. connectivity:

<u>all</u> graphs ε-close to connected in this model

$\Rightarrow$ trivial tester outputs "PASS" on <u>all</u> inputs satisfies requirements of tester

Graph type	max degree	natural representation	notion of distance (changed edges)
sparse	Δ	adjacency list	$\leq \varepsilon \cdot \Delta \cdot n$
dense	n	adjacency matrix	$\leq \varepsilon \cdot n^2$

easier to detect?



Bipartiteness definitions

- can color nodes red/blue s.t. no edge monochromatic

$\equiv$ equivalent definitions!

- can partition nodes into (V_1, V_2) s.t.

$$\underbrace{\nexists\, \underset{(u,v)}{\overset{\shortparallel}{e}} \in E \quad \text{s.t.} \quad \begin{matrix} u,v \in V_1 \\ \text{or } u,v \in V_2 \end{matrix}}_{\text{not bipartite} \Rightarrow \forall (V_1,V_2)\ \exists \text{ "violating edge"}} \Big\}\ \text{"violating edge"}$$

ε-far from bipartite:

- must remove $> \varepsilon n^2$ edges to make bipartite

$\equiv$

- $\forall$ partitions (V_1, V_2) at least εn^2 violating edges

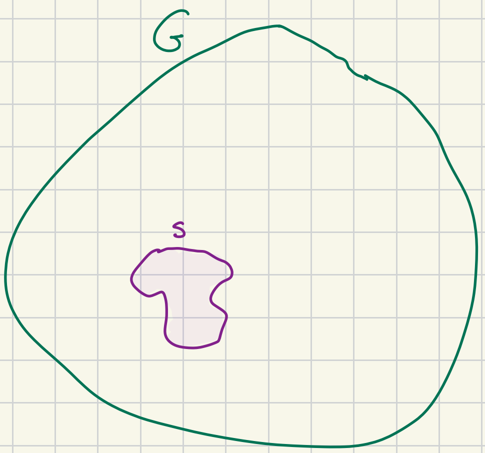
Testing Algorithms

- Test exact bipartiteness: e.g. BFS $\Theta(n)$
- Sparse graph testing $\Theta(\sqrt{n})$
- Proposed property testing algorithm (dense graphs):

  - pick $O(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon})$ size sample of nodes S
  - consider induced graph on S ← only edges between $u,v \in S$

    if bipartite output PASS ← e.g. BFS on tiny graph

    else output FAIL



Actually works!!

A first attempt at a proof?

if G bipartite, induced graph on S bipartite, so algorithm PASSes ✓

if G ε-far from bipartite,

must remove $> \varepsilon n^2$ edges to make it bipartite

equivalently:

$\forall$ partition V_1, V_2 have $> \varepsilon n^2$ violating edges

($> \varepsilon$ fraction of adj matrix slots)

$\Rightarrow$ $\forall$ (V_1, V_2) a sample of edges of size $\geq \Theta(\frac{1}{\varepsilon} \log \frac{1}{\delta})$

hits a viol. edge for V_1, V_2

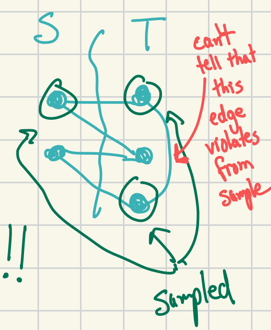
with prob $\geq 1 - (1-\varepsilon)^{\frac{c}{\varepsilon} \log \frac{1}{\delta}} \geq 1 - e^{-c \log \frac{1}{\delta}} = 1 - \delta$

pick # samples s.t. $c = 1$

Great!!! we hit violating edges!!!

Issue #1: how do you know that you will see the violation?

note: no single edge violates all partitions!!



Lets try to use "partition" defn of bipartiteness:

- horrible time complexity
- but maybe query complexity ok?

Algorithm 0

Pick $m = \Theta(?)$ random edge slots & query

$\forall$ partitions (V_1, V_2) of V:

$\text{violating}_{(V_1,V_2)} \leftarrow$ # violating edges in sample wrt V_1, V_2

If $\forall (V_1, V_2)$ $\text{violating}_{(V_1,V_2)} > 0$ then output FAIL

else output PASS

If G bipartite:

bipartition (V_1, V_2) will have $\text{violating}_{(V_1,V_2)} = 0$

$\Rightarrow$ output PASS

If G ε-far from bipartite:

will fail if see violations for all partitions.

how many samples do we need??

Issue #2:

Wait! how small should δ be?

Recall all partitions are bad

But if any partition "looks" good, the algorithm outputs PASS

Pr[any partition "looks" good]: i.e. avoids violating edge

for one partition (V_1, V_2), $\Pr[(V_1, V_2)$ looks good$] \le \delta$

for all partitions (V_1, V_2), $\Pr[$any (V_1, V_2) looks good$] \le \underbrace{2^n}_{\text{union bnd over } 2^n \text{ partitions}} \cdot \delta$

$\Rightarrow$ need $\delta \ll 2^{-n}$

so need # samples of nodes

$$S = \Theta\left(\frac{1}{\varepsilon} \log \frac{1}{\delta}\right) = \Theta\left(\frac{1}{\varepsilon} \log 2^n\right) = \Theta\left(\frac{n}{\varepsilon}\right)$$

query edges $= O(S^2) = \Theta(n^2/\varepsilon^2)$ not sublinear

do we need union bnd?

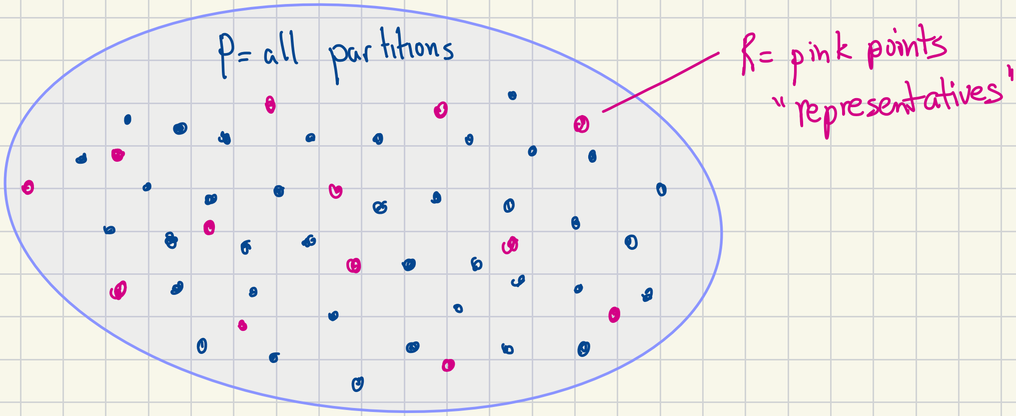
do we need to check all partitions?

idea find (fewer) "representative" partitions

Approach for issue #2:

Plan Consider small set of representative partitions

$|P| = 2^n$



Useful R:

- $R \subseteq P$
- $|R| << |P|$ ← allows to union bound over much smaller set
- every $p \in P$ close to some $r \in R$:

$$\forall p \in P,\ \exists r \in R \text{ s.t. } \mathrm{dist}(p, r) \leq \varepsilon'$$

this property allows us to show that smaller set is still "meaningful"

What does this give us:

(1) if $p \in P$ is bipartition, $\exists r \in R$ with few ($\leq \varepsilon'$) violations

(2) if $\forall p \in P$, p is ε-far from bipartition then, since $R \subseteq P$,

$\forall r \in R$, p is ε-far from bipartition

Plan (continued)

Find "representative" partitions R s.t. all partitions in P are $\frac{\varepsilon}{2}$-close to some representative in $R \subseteq P$

- if G ε-far from bipartite then $\forall$ partitions there are $> \varepsilon n^2$ violating edges

    $\Rightarrow$ $\forall$ representative partitions, have $> \varepsilon n^2$ violating edges (since $R \subseteq P$)

- if G bipartite then $\exists$ partition with 0 violating edges

    $\Rightarrow$ $\exists$ representative partition with $< 0 + \frac{\varepsilon}{2} n^2$ $= \frac{\varepsilon}{2} n^2$ violating edges

Approach for issue #2: test the partition!

Algorithm 1

1. Pick U, U' randomly from V

$\Theta(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ nodes

use to define R, set of partitions

$\Theta(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon})$ nodes

use to test partitions
think of $U' = \{u_1 v_1\ u_2 v_2 \dots\}$
$W = \{(u_1,v_1)\ (u_2,v_2)\dots\}$ pairs

if U not bipartite, FAIL ← $O(|U|)$ queries

2. $\forall$ bipartitions of U into U_1, U_2:

$\leq 2^{|U|}$ of these but not 2^n

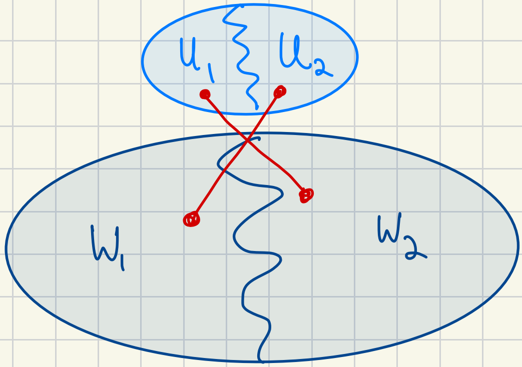
- define oracle (see below) which partitions graph into Z_1, Z_2 based on U_1, U_2
- $\forall u \in U'$ call oracle to see if $u \in Z_1$ or Z_2
- count $\#\{(u,v) \in W$ violating $Z_1, Z_2\}$

  if $\leq \frac{3}{4}\varepsilon$ fraction output PASS

  else continue to next partition

why pass if >0 violations?
note we don't check all partitions.

3. FAIL



Given partition of U into U_1, U_2, define ORACLE to partition whole graph:

Query: node v

Oracle answer: Z_1 or Z_2 or "bad partition"

Oracle algorithm:

output Z_1 if

$v \in U_1$

v has nbr in U_2 but not in U_1

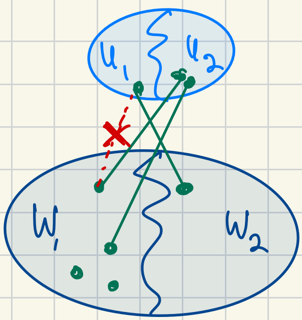
v has no nbr in U_1 or U_2

else output Z_2 if

$v \in U_2$

v has nbr in U_1 but not U_2

else output "bad partition" ← only get here if have nbr to both U_1 & U_2



$Z_1 = U_1 \cup W_1$

$Z_2 = U_2 \cup W_2$

oracle runtime $= O(|U|)$ per query

algorithm 1 queries $= O\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon} \cdot \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ (can reuse same queries for each partition)

algorithm 1 runtime $= O\left(2^{\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}} \times \frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ (since need to try every partition)

no dependence on n

can improve dependence on ε

Behavior

need to show that if G bipartite, likely to pass
+ if G ε-far from bipartite, likely to fail

if G ε-far from bipartite: does it fail?

- <u>all</u> partitions Z_1, Z_2 including those tested by algorithm have $> \varepsilon n^2$ violating edges
- $\forall Z_1, Z_2$ $\Pr[\text{fraction of violating edges in } W \le \frac{3}{4}\varepsilon] << \frac{1}{8 \cdot 2^{|U'|}}$ "test" set from U'

  (chernoff bnd)
- $\Pr[\text{Pass}] = \Pr[\text{any } Z_1 Z_2 \text{ passes}]$

$$\le \underbrace{2^{|U|}}_{\text{union bound over much smaller set}} \cdot \frac{1}{8 \cdot 2^{|U'|}} << \frac{1}{8}$$

(since $|U'| >> |U|$)

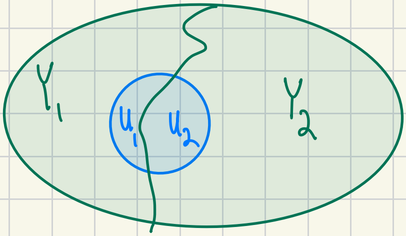
If G bipartite: does it pass?

Let (Y_1, Y_2) be bipartite partition

$\#$violating edges $= 0$

Given sample U,

partition according to Y_1, Y_2



$U_1 \leftarrow U \cap Y_1$

$U_2 \leftarrow U \cap Y_2$

Use (U_1, U_2) to define partition of V into $Q_1^{(U_1,U_2)}, Q_2^{(U_1,U_2)}$

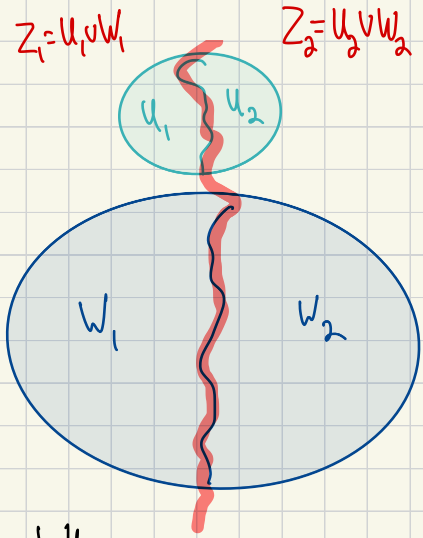
don't actually compute it, but can access it (similar to LCA)

Question how similar is (Y_1, Y_2) to $\underbrace{(Q_1^{U_1,U_2}, Q_2^{U_1,U_2})}_{\text{get from oracle to be defined soon}}$

$\#$ violating edges

Given partition of U into U_1, U_2,

define Oracle to partition whole graph:



Query node v

Oracle answer Z_1, Z_2 or "bad partition"

Oracle algorithm

output Z_1 if

$v \in U_1$

v has a nbr in U_2 but not U_1

only place where Y_1 + Y_2 can differ

v has no nbr in U

else output Z_2 if

$v \in U_2$

v has nbr in U_1 but not U_2

else output "bad partition"

Runtime: $O(|U|)$ per query

$\#$ violating edges in $\left(Q_1^{(u_1,u_2)},\ Q_2^{(u_1,u_2)}\right)$

$\leq \underbrace{0}_{\substack{\#\text{ violating}\\ \text{edges in}\\ (Y_1,Y_2)}} + \underbrace{\#\text{ edges adjacent to any } v \text{ that has no nbr in } \mathcal{U}}$

divide into 2 groups:

$A = \{v \text{ s.t. } \deg(v) < \frac{\varepsilon}{4} n\}$ "small degree"

$B = V \setminus A$ "high degree"

$\leq \underbrace{\frac{\varepsilon}{4} n}_{\substack{\text{max deg}\\ \text{in } A}} \cdot \underbrace{n}_{\substack{\text{upper}\\ \text{bound}\\ \text{on size}\\ \text{of } A}} + \underbrace{n}_{\substack{\text{max}\\ \text{degree}\\ \text{in } B}} \cdot \underbrace{\boxed{}}_{\substack{\text{upper bnd}\\ \text{on size}\\ \text{of } B}}$ see below for $\varepsilon n / 4$

$\circledast$

recall: U is random sample of size $\Theta(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$

$B_U = \{ v$ s.t. $\deg(v) \geq \frac{\varepsilon}{4} n$ & v has no nbr in $U\}$

Lemma $\Pr_U\left[|B_U| \leq \frac{\varepsilon}{4} n \right] \geq 7/8$

Proof $\forall$ v of $\deg \geq \frac{\varepsilon}{4} n$ set $\sigma_v \leftarrow \begin{cases} 1 & \text{if } v \in B_U \\ 0 & \text{o.w.} \end{cases}$

$$
\begin{aligned}
E[\sigma_v] &= \Pr[\sigma_v = 1] \\
&= \left(\Pr[i^{th} \text{ node of } U \text{ isn't nbr of } v] \right)^{|U|} \\
&\leq \Big(\underbrace{1 - \tfrac{\varepsilon}{4}}_{\text{by high degree of } v}\Big)^{|U|} \\
&\leq (1 - \varepsilon/4)^{\Theta(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})} \leq \varepsilon/32
\end{aligned}
$$

$E\Big[\sum_{\substack{v \text{ st.} \\ \deg(v) \geq \frac{\varepsilon}{4} n}} \sigma_v\Big] \leq \frac{\varepsilon n}{32}$ so $\Pr\Big[\sum \sigma_v \geq \underbrace{\frac{8\varepsilon n}{32}}_{\frac{\varepsilon n}{4}}\Big] \leq \frac{1}{8}$ by Markov's $\ddagger$

$\blacksquare$

recap:

$\#$ violating edges in $\left(Q_1^{(u_1,u_2)}, Q_2^{(u_1,u_2)}\right)$

$$\leq \underbrace{0}_{\substack{\#\text{ violating} \\ \text{edges in} \\ (Y_1, Y_2)}} + \underbrace{\#\text{ edges adjacent to any } v \text{ that has no nbr in } \mathcal{U}}_{\text{divide into 2 groups:}}$$

$A = \{v \text{ s.t. } \deg(v) < \frac{\varepsilon}{4} n\}$ "small degree"

$B = V \setminus A$ "high degree"

$$\leq \underbrace{\frac{\varepsilon}{4} n}_{\substack{\text{max deg} \\ \text{in } A}} \cdot \underbrace{n}_{\substack{\text{upper} \\ \text{bound} \\ \text{on size} \\ \text{of } A}} + \underbrace{n}_{\substack{\text{max} \\ \text{degree} \\ \text{in } B}} \cdot \underbrace{\frac{\varepsilon n}{4}}_{\substack{\text{upper bnd} \\ \text{on size} \\ \text{of } B \\ (\text{with prob} \geq 7/8)}} \text{ (from above)} \leq \frac{\varepsilon n^2}{2}$$

$\Rightarrow E[\text{fraction of edges in } Q \text{ violating } W_1^{(u_1,u_2)}, W_2^{(u_1,u_2)}] \leq \varepsilon/2$

So (using Chernoff + samples)

$\Pr[\text{fraction of edges in } Q \text{ violating } W_1^{(u_1,u_2)}, W_2^{(u_1,u_2)} \geq \frac{3\varepsilon}{4}] << 1/8$

$$\Longrightarrow \Pr[\text{output fail}]$$

$$\leq \underbrace{\Pr_u\left[\text{output fail} \;\middle|\; |B_u| > \tfrac{\varepsilon}{4} n\right]}_{\leq 1} \cdot \underbrace{\Pr_u\left[|B_u| > \tfrac{\varepsilon}{4} n\right]}_{\leq 1/8}$$

$$+ \underbrace{\Pr_u\left[\text{output fail} \;\middle|\; |B_u| \leq \tfrac{\varepsilon}{4} n\right]}_{\leq 1/8} \cdot \underbrace{\Pr_u\left[|B_u| \leq \tfrac{\varepsilon}{4} n\right]}_{\leq 1}$$

$$\leq \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$\blacksquare$

<u>Comment</u> can improve runtime to $\mathrm{poly}(1/\varepsilon)$