

Graph Sparsification while Maintaining Cuts

Ramesh Hariharan
Strand Life Sciences

26 Nov 2008

The Setting

- A network G of n nodes and m edges.
- Unweighted for now, we'll handle the weighted case later.
- $m \gg n \log n$
- The goal is to obtain a graph G' which has fewer edges but maintains all the cuts of G approximately.
- G' could be weighted.

The Setting

- A network G of n nodes and m edges.
- Unweighted for now, we'll handle the weighted case later.
- $m \gg n \log n$
- The goal is to obtain a graph G' which has fewer edges but maintains all the cuts of G approximately.
- G' could be weighted.

The Setting

- A network G of n nodes and m edges.
- Unweighted for now, we'll handle the weighted case later.
- $m \gg n \log n$
- The goal is to obtain a graph G' which has fewer edges but maintains all the cuts of G approximately.
- G' could be weighted.

The Setting

- A network G of n nodes and m edges.
- Unweighted for now, we'll handle the weighted case later.
- $m \gg n \log n$
- The goal is to obtain a graph G' which has fewer edges but maintains all the cuts of G approximately.
- G' could be weighted.

The Setting

- A network G of n nodes and m edges.
- Unweighted for now, we'll handle the weighted case later.
- $m \gg n \log n$
- The goal is to obtain a graph G' which has fewer edges but maintains all the cuts of G approximately.
- G' could be weighted.

An Example

- Two nodes with a m edges connected the two.
- Replace by a single edge of weight m .
- The general case is more complex because there are many many cuts in a graph.

An Example

- Two nodes with a m edges connected the two.
- Replace by a single edge of weight m .
- The general case is more complex because there are many many cuts in a graph.

An Example

- Two nodes with a m edges connected the two.
- Replace by a single edge of weight m .
- The general case is more complex because there are many many cuts in a graph.

A Randomized Approach

- Sample each edge e with probability p_e and give it weight $1/p_e$.
- For any cut, its expected weight in the new graph G' is its cardinality in G .
- Question 1: How do we convert this to a high probability bound, i.e., with probability at least half, ALL cuts have weight in G' close to their respective cardinalities in G ?
- Question 2: And how many edges do we have in G' ?

A Randomized Approach

- Sample each edge e with probability p_e and give it weight $1/p_e$.
- For any cut, its expected weight in the new graph G' is its cardinality in G .
- Question 1: How do we convert this to a high probability bound, i.e., with probability at least half, ALL cuts have weight in G' close to their respective cardinalities in G ?
- Question 2: And how many edges do we have in G' ?

A Randomized Approach

- Sample each edge e with probability p_e and give it weight $1/p_e$.
- For any cut, its expected weight in the new graph G' is its cardinality in G .
- Question 1: How do we convert this to a high probability bound, i.e., with probability at least half, ALL cuts have weight in G' close to their respective cardinalities in G ?
- Question 2: And how many edges do we have in G' ?

A Randomized Approach

- Sample each edge e with probability p_e and give it weight $1/p_e$.
- For any cut, its expected weight in the new graph G' is its cardinality in G .
- Question 1: How do we convert this to a high probability bound, i.e., with probability at least half, ALL cuts have weight in G' close to their respective cardinalities in G ?
- Question 2: And how many edges do we have in G' ?

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

Question 2: Determining p_e 's to control size of G'

- Take edges in the min-cut, say cardinality c . Then $p_e = 1/c$ picks one of these c edges in an expected sense.
- Now remove these edges and look at the min-cuts in the resulting connected components, recursively. Two cases for each connected component: min-cut c' now is smaller than c , or larger than c .
- If $c' > c$, then reduce the sampling probability, otherwise, you get too many edges in G' ; sample with probability $1/c'$
- If $c' < c$ retain probability at $1/c$
- Claim 1: The expected number of edges in G' is $(n - 1)$.
- We will need to amplify sampling probabilities by a $\Theta(d \log n)$ factor later; the expected number of edges is then $O(n \log n)$.

A key Property of p_e

- Define k_e to be $1/p_e$ (*edge strength*).
- Claim 2: Any cut in G which contains an edge with strength k has at least k edges with strength k or higher.
- Claim 2': $\sum_{e \in \text{cut} X} p_e \geq 1$

A key Property of p_e

- Define k_e to be $1/p_e$ (*edge strength*).
- Claim 2: Any cut in G which contains an edge with strength k has at least k edges with strength k or higher.
- Claim 2': $\sum_{e \in \text{cut} X} p_e \geq 1$

A key Property of p_e

- Define k_e to be $1/p_e$ (*edge strength*).
- Claim 2: Any cut in G which contains an edge with strength k has at least k edges with strength k or higher.
- Claim 2': $\sum_{e \in \text{cut}X} p_e \geq 1$

Chernoff Bound

- Suppose we sample each edge e independently with probability $\min(1, \Delta * p_e)$ as above and give all edges the same weight 1.
- The probability that cut X in G' deviates from this expectation by more than $\epsilon E(X)$ is at most $e^{-\Theta(\epsilon^2 E(X))}$
- We want that $\sum_X e^{-\Theta(\epsilon^2 E(X))} = \sum_X e^{-\Theta(\sum_{e \in X} \min(\epsilon^2, \epsilon^2 \Delta p_e))}$ be smaller than, say $1/2$.
- Set $\Delta = \Theta(d \log n / \epsilon^2)$; we have $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- The term for X starts looking like $1/n^d$ by Claim 2'; does this converge when added over all cuts X ?

Chernoff Bound

- Suppose we sample each edge e independently with probability $\min(1, \Delta * p_e)$ as above and give all edges the same weight 1.
- The probability that cut X in G' deviates from this expectation by more than $\epsilon E(X)$ is at most $e^{-\Theta(\epsilon^2 E(X))}$
- We want that $\sum_X e^{-\Theta(\epsilon^2 E(X))} = \sum_X e^{-\Theta(\sum_{e \in X} \min(\epsilon^2, \epsilon^2 \Delta p_e))}$ be smaller than, say $1/2$.
- Set $\Delta = \Theta(d \log n / \epsilon^2)$; we have $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- The term for X starts looking like $1/n^d$ by Claim 2'; does this converge when added over all cuts X ?

Chernoff Bound

- Suppose we sample each edge e independently with probability $\min(1, \Delta * p_e)$ as above and give all edges the same weight 1.
- The probability that cut X in G' deviates from this expectation by more than $\epsilon E(X)$ is at most $e^{-\Theta(\epsilon^2 E(X))}$
- We want that $\sum_X e^{-\Theta(\epsilon^2 E(X))} = \sum_X e^{-\Theta(\sum_{e \in X} \min(\epsilon^2, \epsilon^2 \Delta p_e))}$ be smaller than, say $1/2$.
- Set $\Delta = \Theta(d \log n / \epsilon^2)$; we have $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- The term for X starts looking like $1/n^d$ by Claim 2'; does this converge when added over all cuts X ?

Chernoff Bound

- Suppose we sample each edge e independently with probability $\min(1, \Delta * p_e)$ as above and give all edges the same weight 1.
- The probability that cut X in G' deviates from this expectation by more than $\epsilon E(X)$ is at most $e^{-\Theta(\epsilon^2 E(X))}$
- We want that $\sum_X e^{-\Theta(\epsilon^2 E(X))} = \sum_X e^{-\Theta(\sum_{e \in X} \min(\epsilon^2, \epsilon^2 \Delta p_e))}$ be smaller than, say $1/2$.
- Set $\Delta = \Theta(d \log n / \epsilon^2)$; we have $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- The term for X starts looking like $1/n^d$ by Claim 2'; does this converge when added over all cuts X ?

Chernoff Bound

- Suppose we sample each edge e independently with probability $\min(1, \Delta * p_e)$ as above and give all edges the same weight 1.
- The probability that cut X in G' deviates from this expectation by more than $\epsilon E(X)$ is at most $e^{-\Theta(\epsilon^2 E(X))}$
- We want that $\sum_X e^{-\Theta(\epsilon^2 E(X))} = \sum_X e^{-\Theta(\sum_{e \in X} \min(\epsilon^2, \epsilon^2 \Delta p_e))}$ be smaller than, say $1/2$.
- Set $\Delta = \Theta(d \log n / \epsilon^2)$; we have $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- The term for X starts looking like $1/n^d$ by Claim 2'; does this converge when added over all cuts X ?

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Convergence of sum over all cuts

- $\sum_X e^{-\sum_{e \in X} \min(\Theta(\epsilon^2), d \log n * p_e)}$.
- In any weighted graph, the number of cuts with weight α times the min-cut is at most $n^{2\alpha}$.
- Consider a graph H , same edges as G , but with edge weights $\min(-\Theta(\epsilon^2), d \log n * p_e)$.
- H has min-cut at least $d \log n$ by Claims 2,2'.
- We need to bound $\sum_{X \in H} e^{-w_H(X)}$, given H has min-cut at least $d \log n$.
- Easily seen to be $O(n^{-d-2})$; set $d > 3$ say.

Handling Weights via Equiweighted Partitions

- Corollary from above: if we sample with probability $p'_e = \min(1, \Delta * p_e)$ and give each edge weight 1 then each cut is within a $1 \pm \epsilon$ factor of its expectation. Same holds if each edge weight is say k .
- But we need to give weights $1/p'_e$ to make expectations equal to the original value; so edge weights are not the same.
- So partition G into a family of graphs F_i so each F_i has edges of the same weight; at most m graphs; if the above corollary holds for each F_i , then it holds for all F_i 's with probability m/n^{d-2} so we can set $d > 4$.
- Does the corollary hold for each F_i ???
- No because claims 2,2' do not hold!

Handling Weights via Equiweighted Partitions

- Corollary from above: if we sample with probability $p'_e = \min(1, \Delta * p_e)$ and give each edge weight 1 then each cut is within a $1 \pm \epsilon$ factor of its expectation. Same holds if each edge weight is say k .
- But we need to give weights $1/p'_e$ to make expectations equal to the original value; so edge weights are not the same.
- So partition G into a family of graphs F_i so each F_i has edges of the same weight; at most m graphs; if the above corollary holds for each F_i , then it holds for all F_i 's with probability m/n^{d-2} so we can set $d > 4$.
- Does the corollary hold for each F_i ??.
- No because claims 2,2' do not hold!

Handling Weights via Equiweighted Partitions

- Corollary from above: if we sample with probability $p'_e = \min(1, \Delta * p_e)$ and give each edge weight 1 then each cut is within a $1 \pm \epsilon$ factor of its expectation. Same holds if each edge weight is say k .
- But we need to give weights $1/p'_e$ to make expectations equal to the original value; so edge weights are not the same.
- So partition G into a family of graphs F_i so each F_i has edges of the same weight; at most m graphs; if the above corollary holds for each F_i , then it holds for all F_i 's with probability m/n^{d-2} so we can set $d > 4$.
- Does the corollary hold for each F_i ???
- No because claims 2,2' do not hold!

Handling Weights via Equiweighted Partitions

- Corollary from above: if we sample with probability $p'_e = \min(1, \Delta * p_e)$ and give each edge weight 1 then each cut is within a $1 \pm \epsilon$ factor of its expectation. Same holds if each edge weight is say k .
- But we need to give weights $1/p'_e$ to make expectations equal to the original value; so edge weights are not the same.
- So partition G into a family of graphs F_i so each F_i has edges of the same weight; at most m graphs; if the above corollary holds for each F_i , then it holds for all F_i 's with probability m/n^{d-2} so we can set $d > 4$.
- Does the corollary hold for each F_i ??.
- No because claims 2,2' do not hold!

Handling Weights via Equiweighted Partitions

- Corollary from above: if we sample with probability $p'_e = \min(1, \Delta * p_e)$ and give each edge weight 1 then each cut is within a $1 \pm \epsilon$ factor of its expectation. Same holds if each edge weight is say k .
- But we need to give weights $1/p'_e$ to make expectations equal to the original value; so edge weights are not the same.
- So partition G into a family of graphs F_i so each F_i has edges of the same weight; at most m graphs; if the above corollary holds for each F_i , then it holds for all F_i 's with probability m/n^{d-2} so we can set $d > 4$.
- Does the corollary hold for each F_i ??.
- No because claims 2,2' do not hold!

A Different Equiweighted Partition

- Add to F_i all edges in $F_{i+1}, F_{i+2}..$ etc.
- The differential weight of edges in F_i is the actual weight minus weight already accounted for by $F_1 \dots F_{i-1}$.
- Claims 2,2' hold!

A Different Equiweighted Partition

- Add to F_i all edges in $F_{i+1}, F_{i+2}..$ etc.
- The differential weight of edges in F_i is the actual weight minus weight already accounted for by $F_1 \dots F_{i-1}$.
- Claims 2,2' hold!

A Different Equiweighted Partition

- Add to F_i all edges in $F_{i+1}, F_{i+2}..$ etc.
- The differential weight of edges in F_i is the actual weight minus weight already accounted for by $F_1 \dots F_{i-1}$.
- Claims 2,2' hold!

Computing p_e

- Computing a value $\geq p_e$ (or equivalently $\leq k_e$) which still ensures that the expected number of edges is small is ok! Why?
- The Nagamochi-Ibaraki Certificate: Build k forests so if two vertices are connected in a particular forest they are connected in all previous forests. Then all edges involved in cuts of size k or smaller are captured in these forests. Time is linear.
- So for any edge e , the index of the last forest in which the endpoints of e are connected to each other is a lower bound on the connectivity between these endpoints.
- But k_e could be much smaller than the above connectivity, so the above does not give a lower bound for this.

Computing p_e

- Computing a value $\geq p_e$ (or equivalently $\leq k_e$) which still ensures that the expected number of edges is small is ok! Why?
- The Nagamochi-Ibaraki Certificate: Build k forests so if two vertices are connected in a particular forest they are connected in all previous forests. Then all edges involved in cuts of size k or smaller are captured in these forests. Time is linear.
- So for any edge e , the index of the last forest in which the endpoints of e are connected to each other is a lower bound on the connectivity between these endpoints.
- But k_e could be much smaller than the above connectivity, so the above does not give a lower bound for this.

Computing p_e

- Computing a value $\geq p_e$ (or equivalently $\leq k_e$) which still ensures that the expected number of edges is small is ok! Why?
- The Nagamochi-Ibaraki Certificate: Build k forests so if two vertices are connected in a particular forest they are connected in all previous forests. Then all edges involved in cuts of size k or smaller are captured in these forests. Time is linear.
- So for any edge e , the index of the last forest in which the endpoints of e are connected to each other is a lower bound on the connectivity between these endpoints.
- But k_e could be much smaller than the above connectivity, so the above does not give a lower bound for this.

Computing p_e

- Computing a value $\geq p_e$ (or equivalently $\leq k_e$) which still ensures that the expected number of edges is small is ok! Why?
- The Nagamochi-Ibaraki Certificate: Build k forests so if two vertices are connected in a particular forest they are connected in all previous forests. Then all edges involved in cuts of size k or smaller are captured in these forests. Time is linear.
- So for any edge e , the index of the last forest in which the endpoints of e are connected to each other is a lower bound on the connectivity between these endpoints.
- But k_e could be much smaller than the above connectivity, so the above does not give a lower bound for this.

Lower bounding k_e

- Run NI for $4k$ forests, remove these edges and repeat.
- After $\log n$ iterations, all remaining edges have $k_e > k$. Why?
- Repeat this for doubling values of k gives the reqd estimates.

Lower bounding k_e

- Run NI for $4k$ forests, remove these edges and repeat.
- After $\log n$ iterations, all remaining edges have $k_e > k$. Why?
- Repeat this for doubling values of k gives the reqd estimates.

An Alternative

- What if p_e is set to reciprocal of the edge connectivity between the endpoints of e ? Note that Nagamochi-Ibaraki gives us lower-bounds to the edge connectivity of e directly.
- Do Claims 2,2' hold?
- For the whole graph? For the F_i 's?

An Alternative

- What if p_e is set to reciprocal of the edge connectivity between the endpoints of e ? Note that Nagamochi-Ibaraki gives us lower-bounds to the edge connectivity of e directly.
- Do Claims 2,2' hold?
- For the whole graph? For the F_i 's?

An Alternative

- What if p_e is set to reciprocal of the edge connectivity between the endpoints of e ? Note that Nagamochi-Ibaraki gives us lower-bounds to the edge connectivity of e directly.
- Do Claims 2,2' hold?
- For the whole graph? For the F_i 's?

What remains?

- Proof of Chernoff bound? Why do we need the decomposition into the F_i 's.
- Running time of Nagamochi-Ibaraki.
- Proof of the bound on the number of near minimum cuts.

What remains?

- Proof of Chernoff bound? Why do we need the decomposition into the F_i 's.
- Running time of Nagamochi-Ibaraki.
- Proof of the bound on the number of near minimum cuts.

What remains?

- Proof of Chernoff bound? Why do we need the decomposition into the F_i 's.
- Running time of Nagamochi-Ibaraki.
- Proof of the bound on the number of near minimum cuts.