KADABRA is an ADaptive Algorithm for Betweenness via Random Approximation

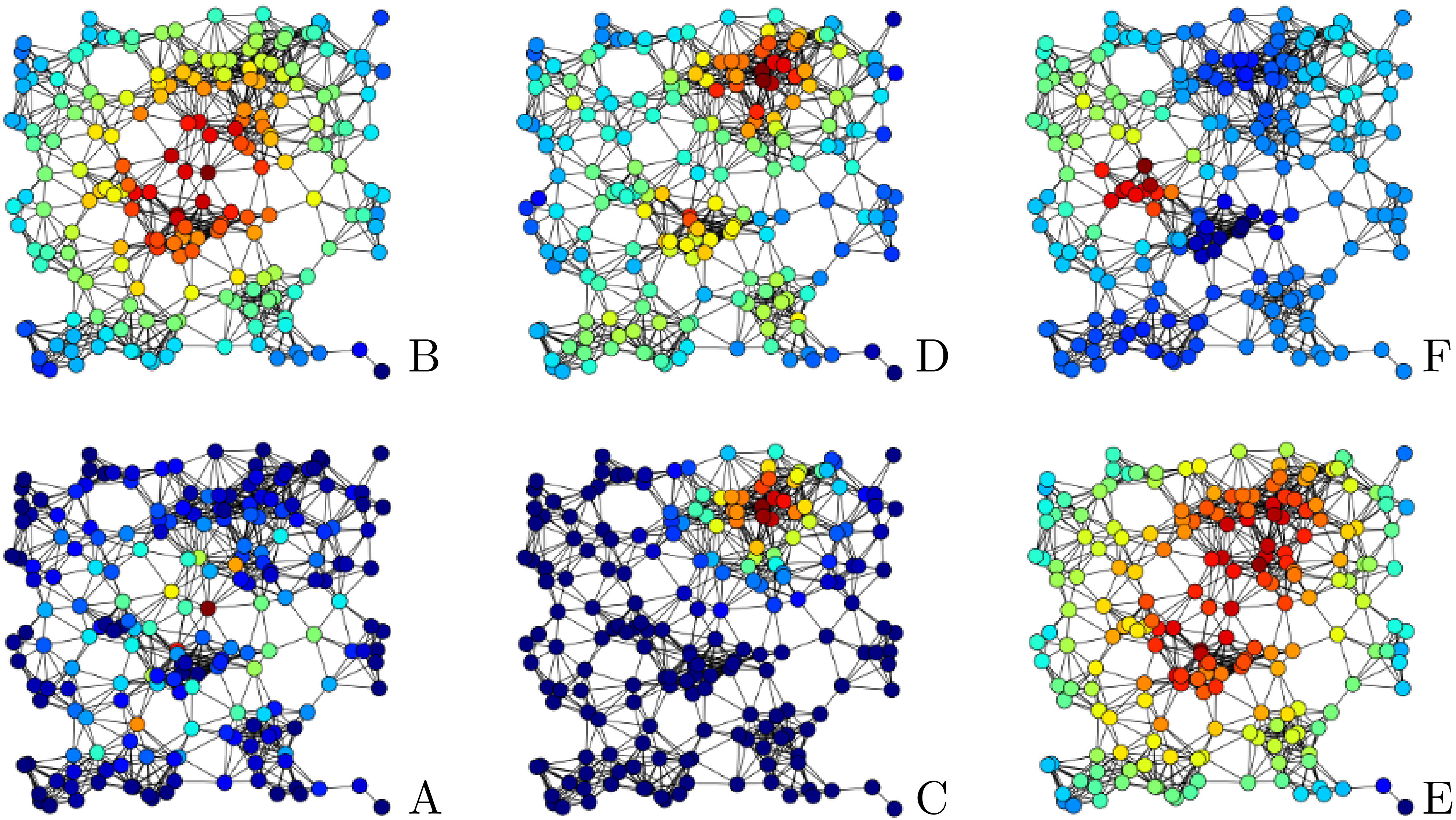
Emanuele Natale[†]
joint work with Michele Borassi[*]

[†] SAPIENZA UNIVERSITÀ DI ROMA

[*] IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

24rd Annual European Symposium on Algorithms
(ESA 2016)
22-26 August, 2016
Aarhus, Denmark

Centrality Measures in (Complex) Networks



Examples of A) Betweenness centrality, B) Closeness centrality, C) Eigenvector centrality, D) Degree centrality, E) Harmonic Centrality and F) Katz centrality of the same graph*.

*Source: Wikipedia.

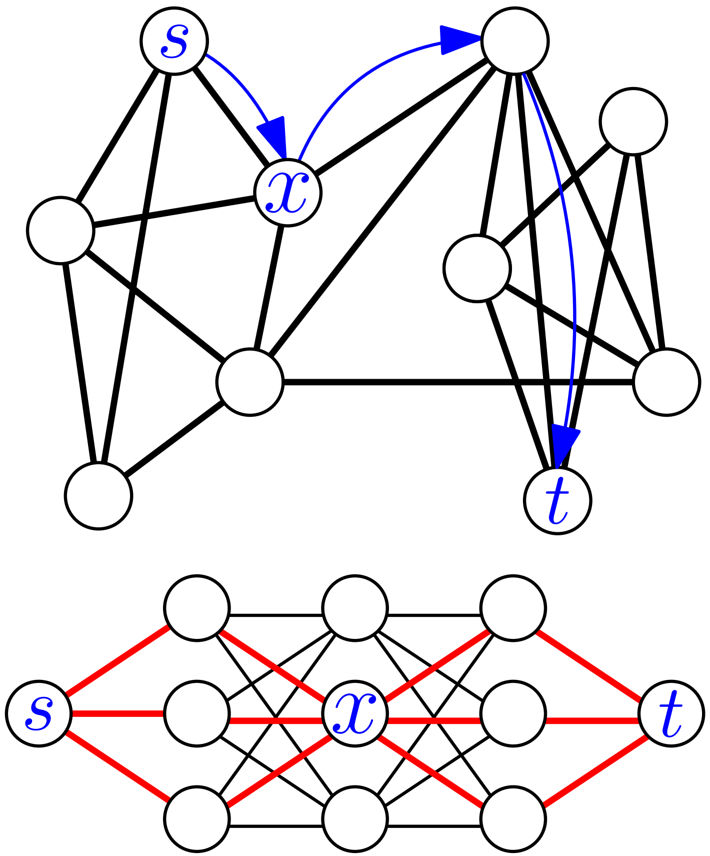
Centrality Measures in (Complex) Networks

Betweenness centrality:

$$\sigma(x) = \frac{1}{n(n-1)} \sum_{s \neq x \neq t} \frac{\sigma_{st}(x)}{\sigma_{st}}$$

σ_{st} := # shortest paths from s to t

$\sigma_{st}(x)$:= # shortest paths from s to t through x



Centrality Measures in (Complex) Networks

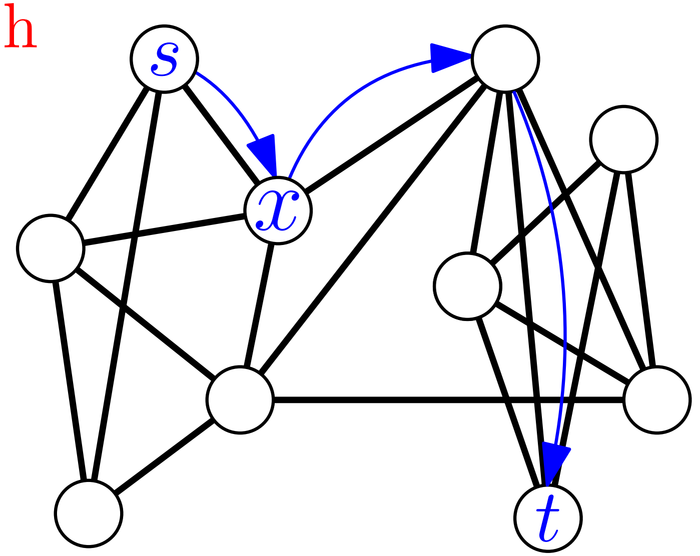
Prob. $\Pr(X)$ of being in a shortest path

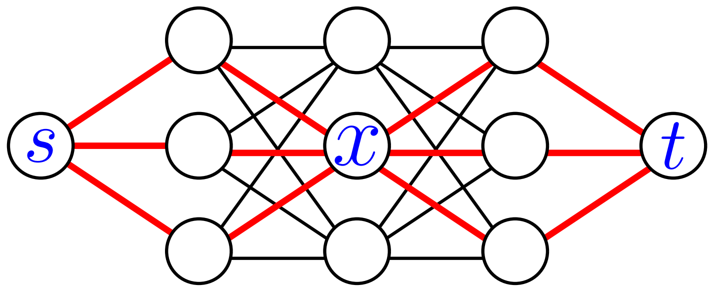
Betweenness centrality:

$$\sigma(x) = \frac{1}{n(n-1)} \sum_{s \neq x \neq t} \frac{\sigma_{st}(x)}{\sigma_{st}}$$

σ_{st} := # shortest paths from s to t

$\sigma_{st}(x)$:= # shortest paths from s to t through x



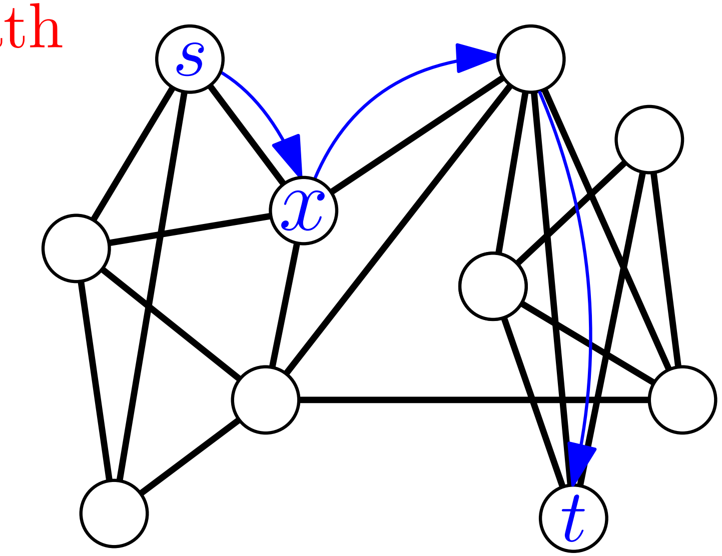


Centrality Measures in (Complex) Networks

Prob. $\Pr(X)$ of being in a shortest path

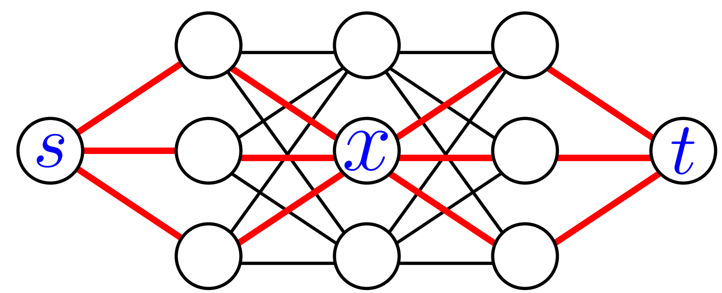
Betweenness centrality:

$$\sigma(x) = \frac{1}{n(n-1)} \sum_{s \neq x \neq t} \frac{\sigma_{st}(x)}{\sigma_{st}}$$



$\sigma_{st} :=$ # shortest paths from s to t

$\sigma_{st}(x) :=$ # shortest paths from s to t through x



[Brandes '01]: betweenness of all nodes in $\mathcal{O}(mn)$

[Borassi et al. '15]: betweenness of a node requires $\Omega(n^{2-\epsilon})$ on sparse graphs (assuming SETH)

Random Approximation of Centrality

Eppstein, Wang [SODA '01]: samples $S \subset V$ and compute measure w.r.t. $S \implies$ approx. of *closeness centrality* w.h.p. in $\mathcal{O}(CB) \cdot \mathcal{O}(SSSP)$

Random Approximation of Centrality

Eppstein, Wang [SODA '01]: samples $S \subset V$ and compute measure w.r.t. $S \implies$ approx. of *closeness centrality* w.h.p. in $\mathcal{O}(CB) \cdot \mathcal{O}(SSSP)$

Betweenness centrality:

Idea: samples $s, t \in V$ and give 1 point to x if x is in the st-shortest path

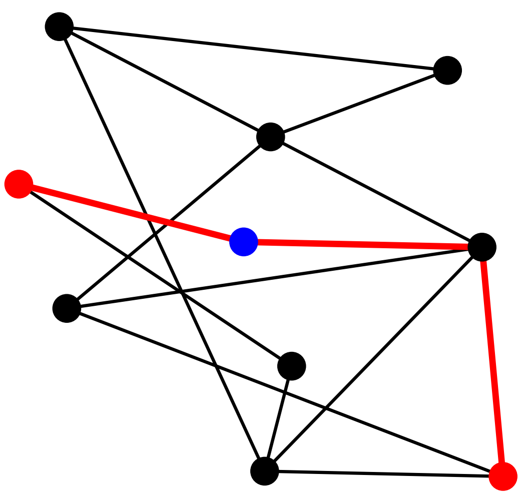
Random Approximation of Centrality

Eppstein, Wang [SODA '01]: samples $S \subset V$ and compute measure w.r.t. $S \implies$ approx. of *closeness centrality* w.h.p. in $\mathcal{O}(CB) \cdot \mathcal{O}(SSSP)$

Betweenness centrality:

Idea: samples $s, t \in V$ and give 1 point to x if x is in the st-shortest path

......, Riondato and Upfal [KDD '16]: ABRA*, ϵ-approx. in time $\mathcal{O}(g(\mathrm{RA})) \cdot \mathcal{O}(st\text{-}SP)$



* Approximating Betweenness via Rademacher Averages

Random Approximation of Centrality

Eppstein, Wang [SODA '01]: samples $S \subset V$ and compute measure w.r.t. $S \implies$ approx. of *closeness centrality* w.h.p. in $\mathcal{O}(CB) \cdot \mathcal{O}(SSSP)$

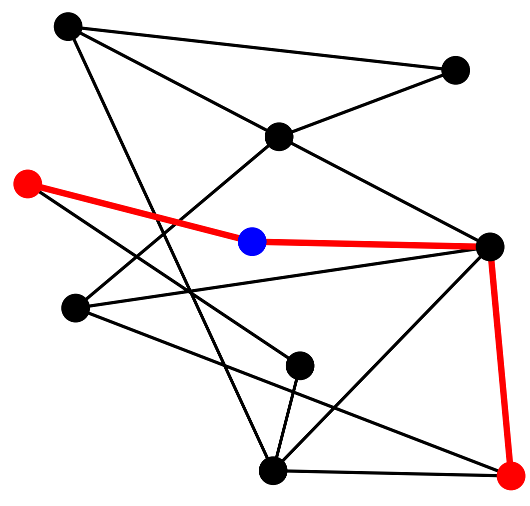
#samples shortest paths

Betweenness centrality:

Idea: samples $s, t \in V$ and give 1 point to x if x is in the st-shortest path

......, Riondato and Upfal [KDD '16]:
ABRA*, ϵ-approx. in time
$\mathcal{O}(g(\text{RA})) \cdot \mathcal{O}(st\text{-}SP)$

#samples shortest paths



* Approximating Betweenness via Rademacher Averages

Random Approximation of Centrality

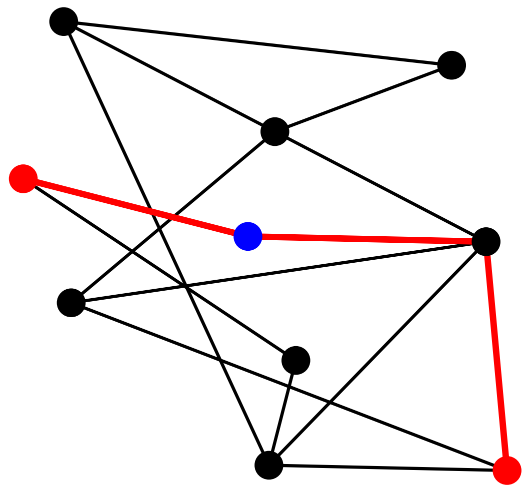
Eppstein, Wang [SODA '01]: samples $S \subset V$ and compute measure w.r.t. $S \implies$ approx. of *closeness centrality* w.h.p. in $\mathcal{O}(CB) \cdot \mathcal{O}(SSSP)$

Betweenness centrality:

Idea: samples $s, t \in V$ and give 1 point to x if x is in the st-shortest path

……., Riondato and Upfal [KDD '16]:
ABRA*, ϵ-approx. in time
$\mathcal{O}(g(\mathrm{RA})) \cdot \mathcal{O}(st\text{-}SP)$

Rademacher Averages.
We could, but we keep it simple.



* Approximating Betweenness via Rademacher Averages

KADABRA: Overview

Input: graph $G = (V, E)$, error prob. δ, error approx. λ

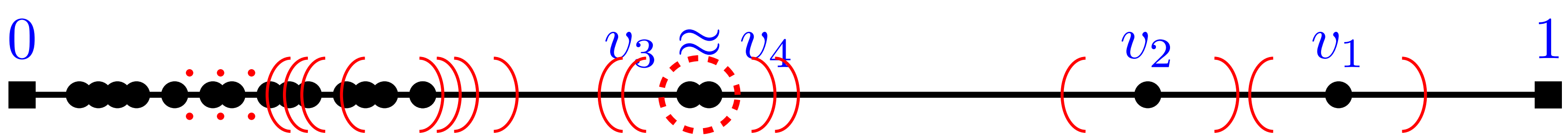
1. $\ldots$ 2. $\ldots$ 3. $\ldots$
4. **foreach** $v \in V$ **do** $\tilde{\mathbf{b}}(v) \leftarrow 0$
5. **while** $(\tau < \omega) \wedge (\neg$**haveToStop**$(\ldots))$
6. $\quad \pi \leftarrow$ **samplePath**$()$
7. $\quad\quad$ **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v) + 1$, $\tau \leftarrow \tau + 1$
8. **end while**
9. **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v)/\tau$
10. **return** $\tilde{\mathbf{b}}$

KADABRA: Overview

Input: graph $G = (V, E)$, error prob. δ, error approx. λ

1. … 2. … 3. …
4. **foreach** $v \in V$ **do** $\tilde{\mathbf{b}}(v) \leftarrow 0$
5. **while** $(\tau < \omega) \wedge (\neg$**haveToStop**$(\dots))$
6. $\quad \pi \leftarrow$ **samplePath**$()$
7. $\quad$ **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v) + 1$, $\tau \leftarrow \tau + 1$
8. **end while**
9. **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v)/\tau$
10. **return** $\tilde{\mathbf{b}}$

TOP-k centralities:

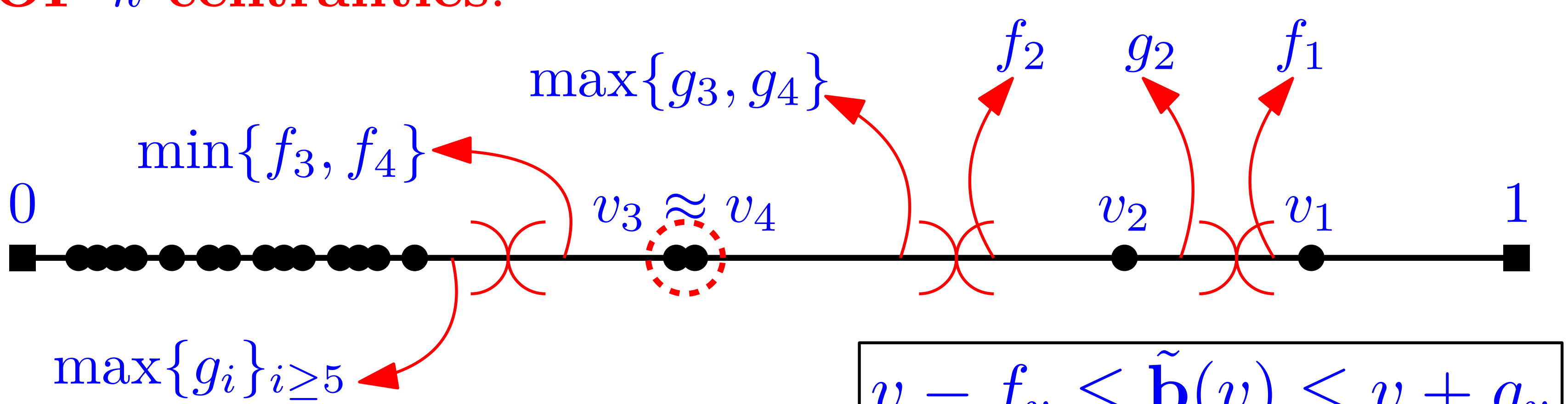


KADABRA: Overview

Input: graph $G = (V, E)$, error prob. δ, error approx. λ

1. $\ldots$ 2. $\ldots$ 3. $\ldots$
4. **foreach** $v \in V$ **do** $\tilde{\mathbf{b}}(v) \leftarrow 0$
5. **while** $(\tau < \omega) \wedge (\neg$**haveToStop**$(\ldots))$
6. $\quad \pi \leftarrow$ **samplePath**$()$
7. $\quad$ **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v) + 1$, $\tau \leftarrow \tau + 1$
8. **end while**
9. **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v)/\tau$
10. **return** $\tilde{\mathbf{b}}$

TOP-k centralities:



$$v - f_v \leq \tilde{\mathbf{b}}(v) \leq v + g_v$$

KADABRA: Overview

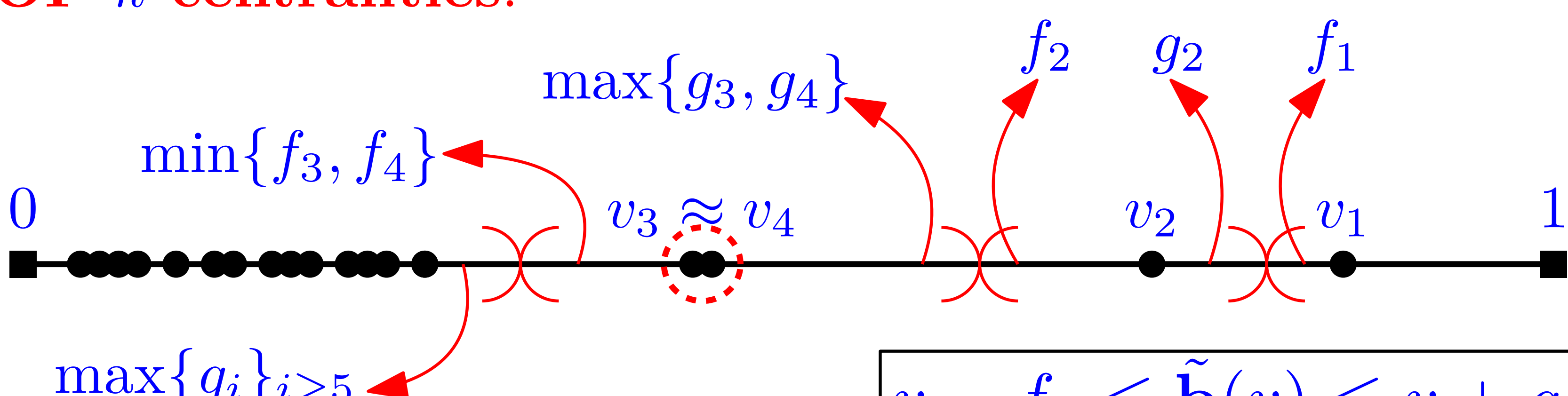
Input: graph $G = (V, E)$, error prob. δ, error approx. λ

1. … 2. … 3. …
4. **foreach** $v \in V$ **do** $\tilde{\mathbf{b}}(v) \leftarrow 0$
5. **while** $(\tau < \omega) \wedge (\neg$ **haveToStop**$(\dots))$
6. $\quad \pi \leftarrow$ **samplePath**$()$
7. $\quad\quad$ **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v) + 1$, $\tau \leftarrow \tau + 1$
8. **end while**
9. **for each** $v \in \pi$ **do** $\tilde{\mathbf{b}}(v) \leftarrow \tilde{\mathbf{b}}(v)/\tau$
10. **return** $\tilde{\mathbf{b}}$

2^{nd} contribution: Adaptive Sampling Made Rigorous

1^{st} contribution: BBBFS

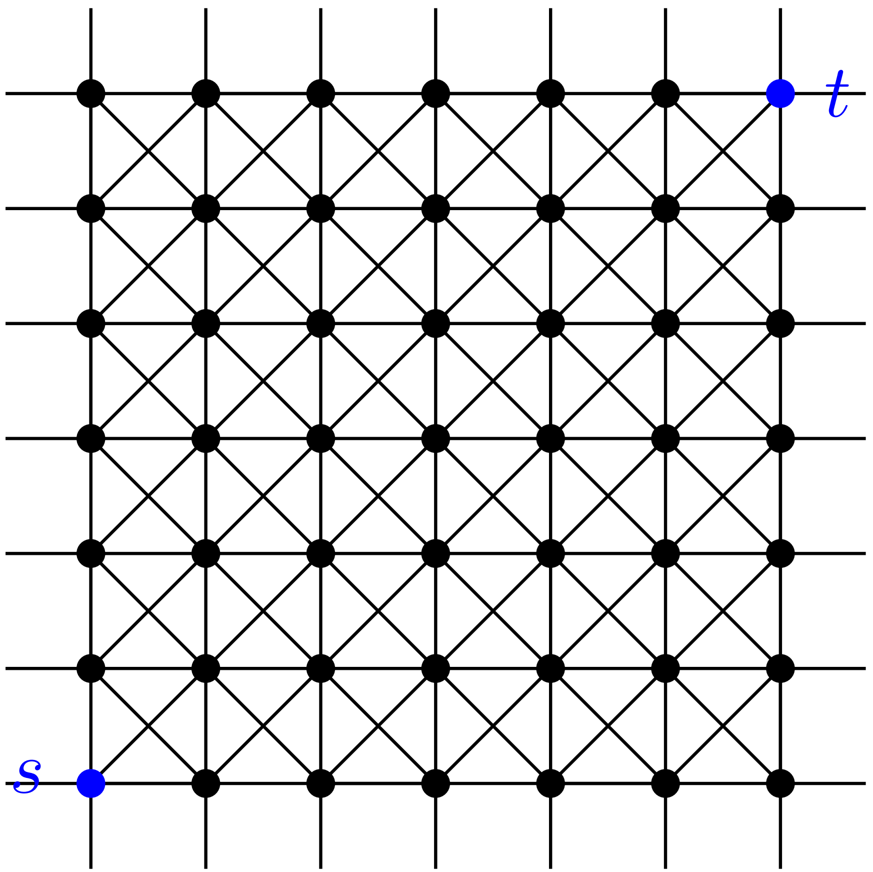
TOP-k centralities:



$$v - f_v \leq \tilde{\mathbf{b}}(v) \leq v + g_v$$

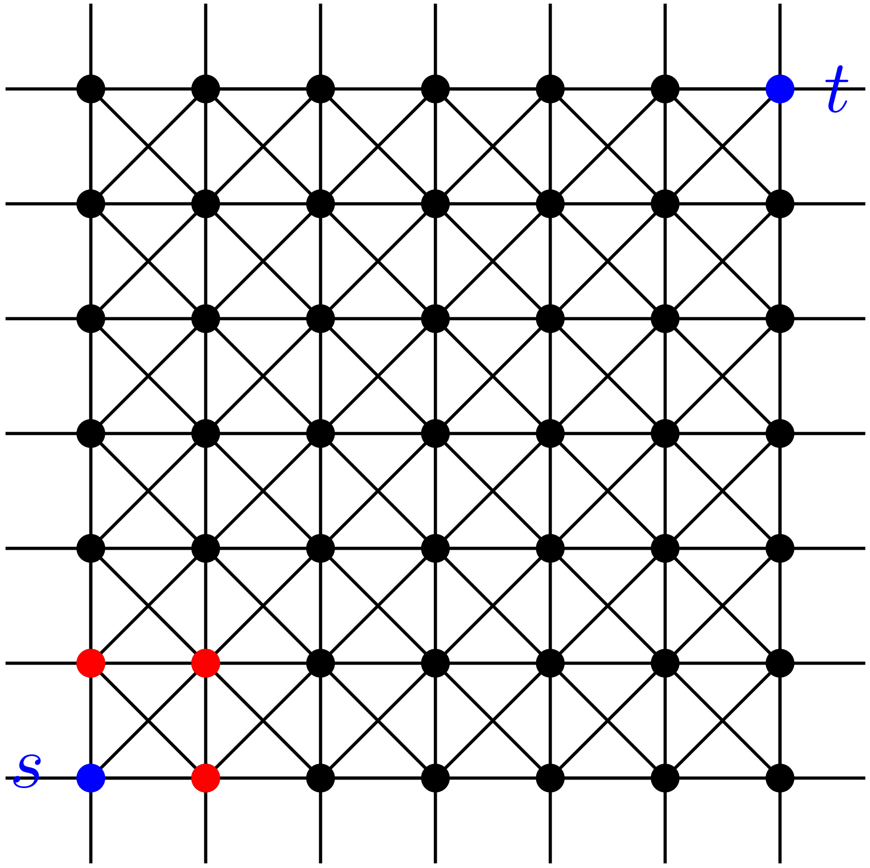
Our Idea: Balanced Bidirectional BFS

Simple BFS



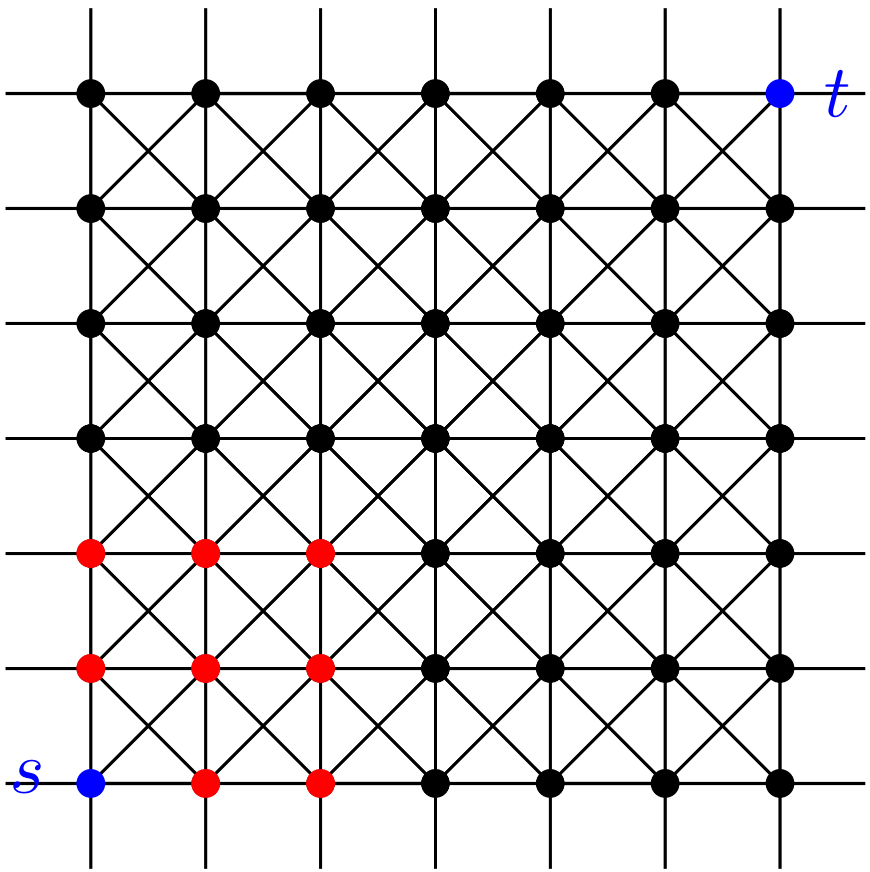
Our Idea: Balanced Bidirectional BFS

Simple BFS



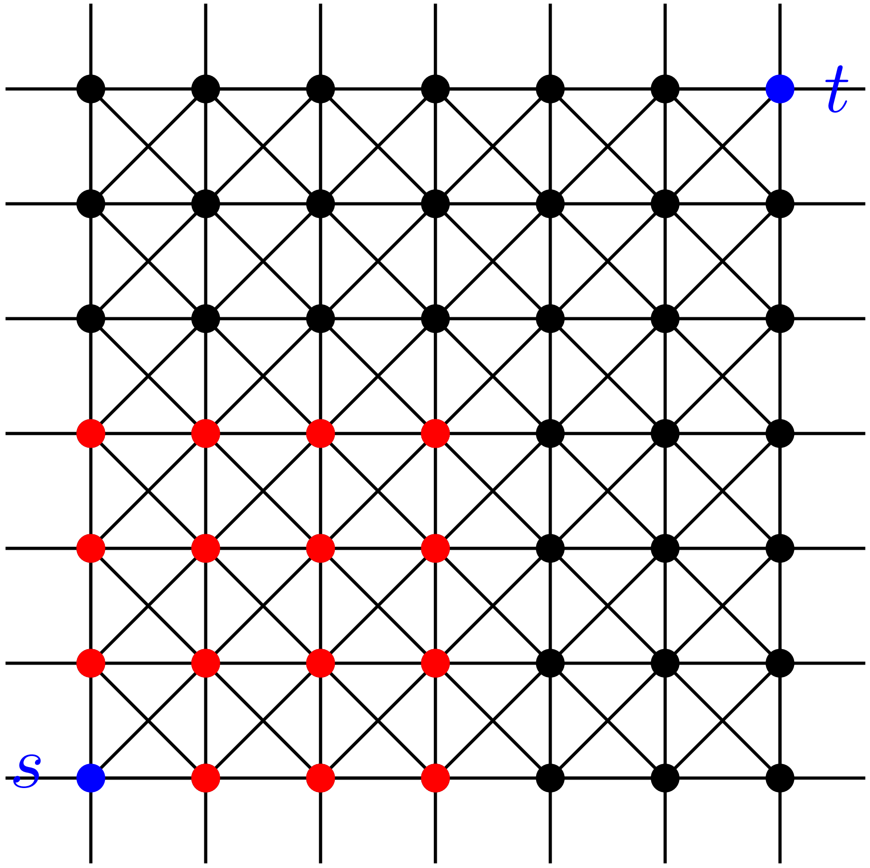
Our Idea: Balanced Bidirectional BFS

Simple BFS



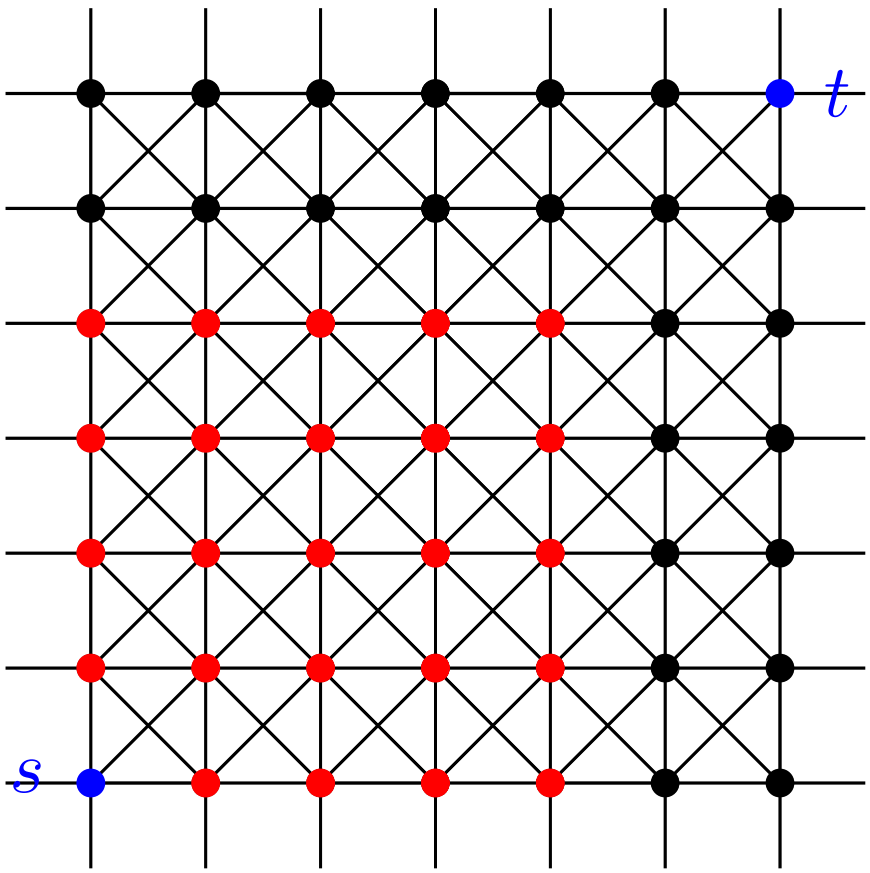
Our Idea: Balanced Bidirectional BFS

Simple BFS



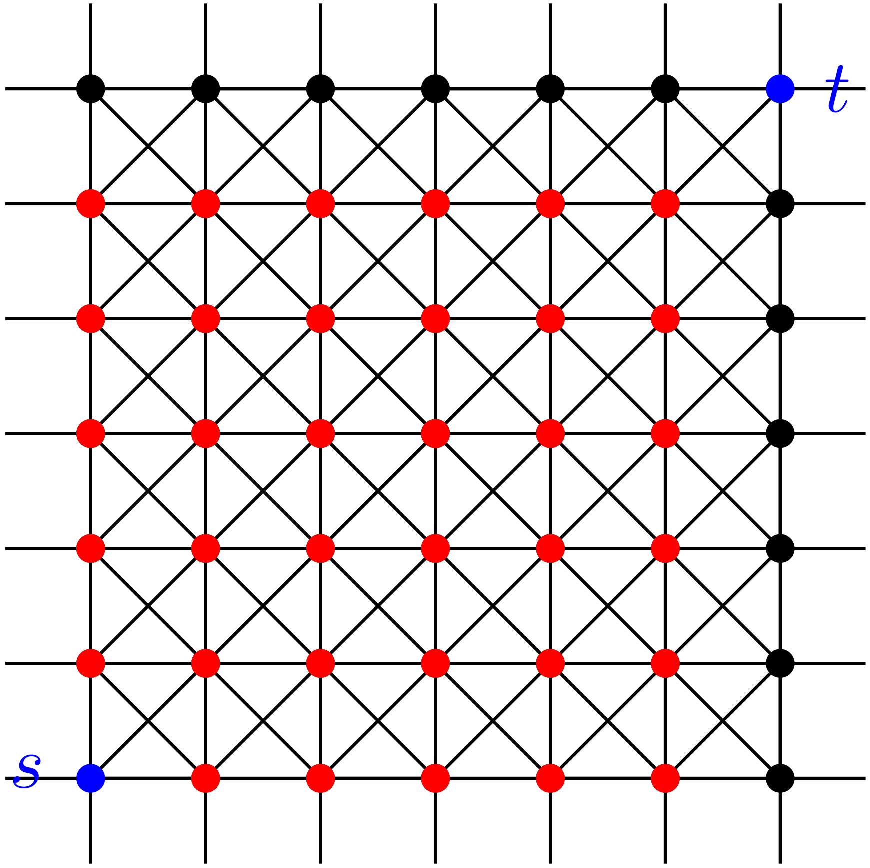
Our Idea: Balanced Bidirectional BFS

Simple BFS



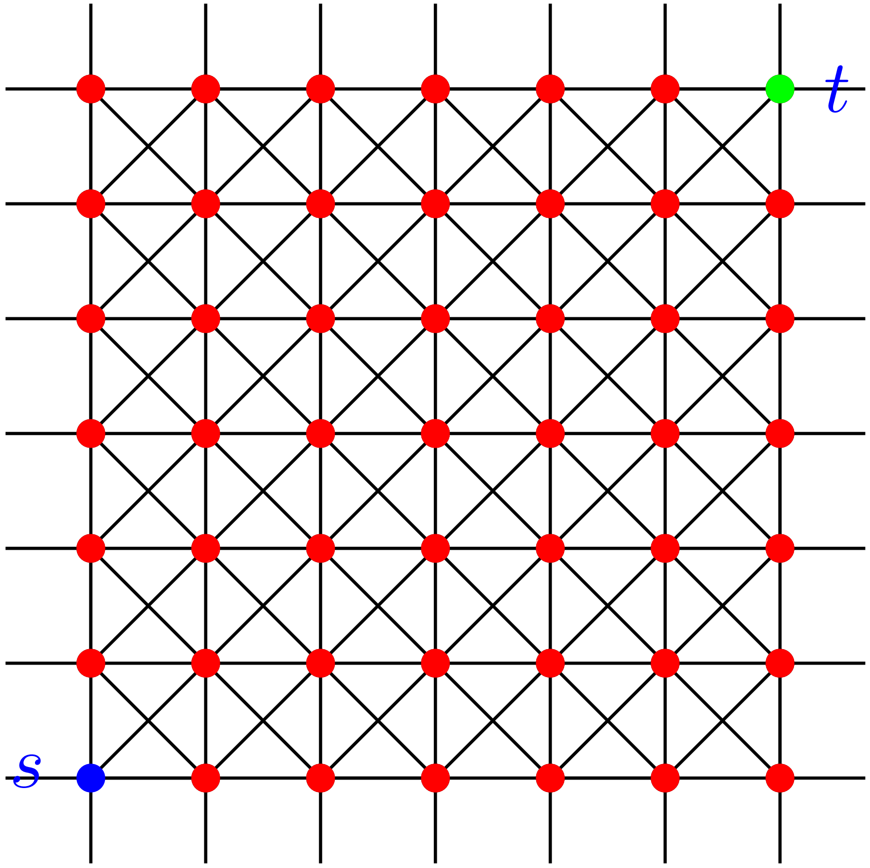
Our Idea: Balanced Bidirectional BFS

Simple BFS



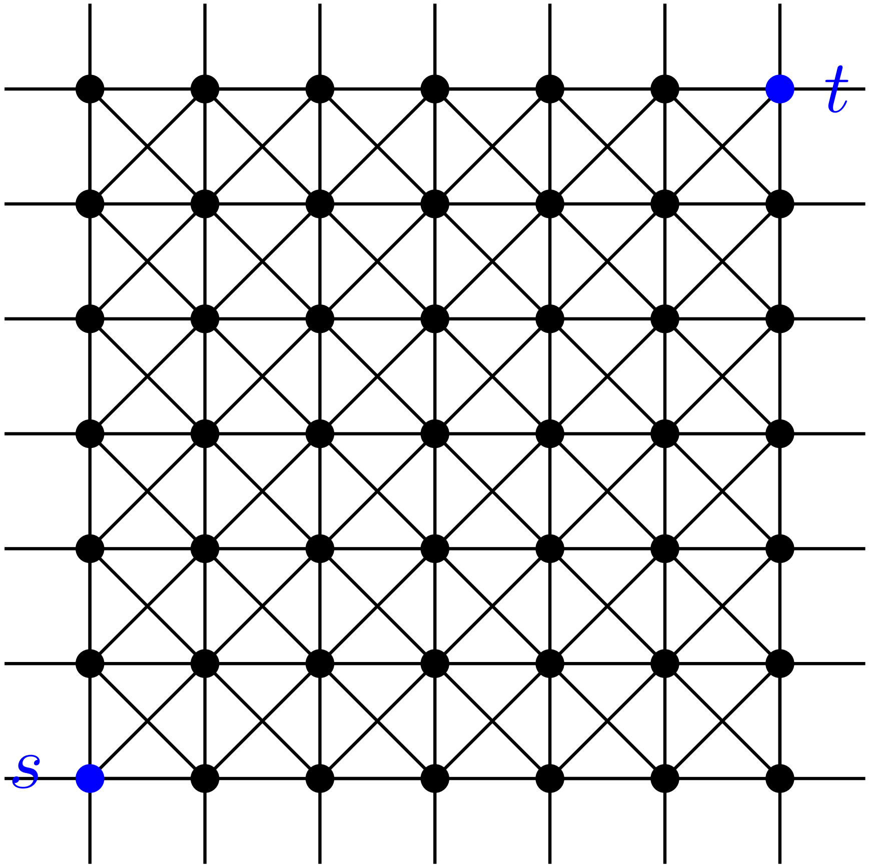
Our Idea: Balanced Bidirectional BFS

Simple BFS



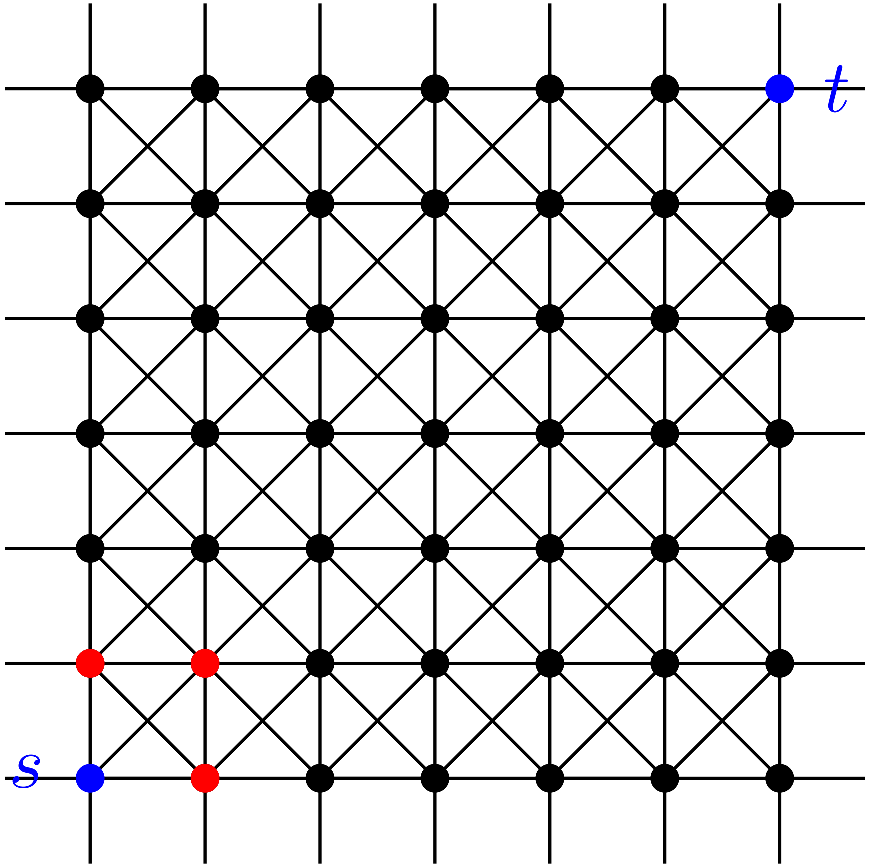
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



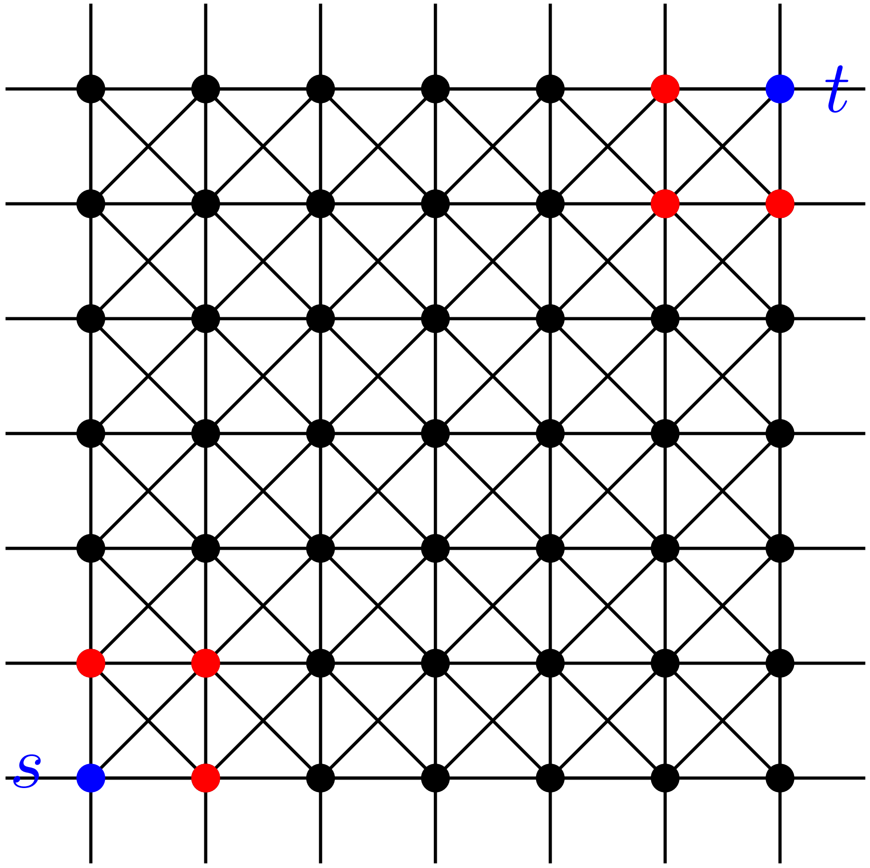
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



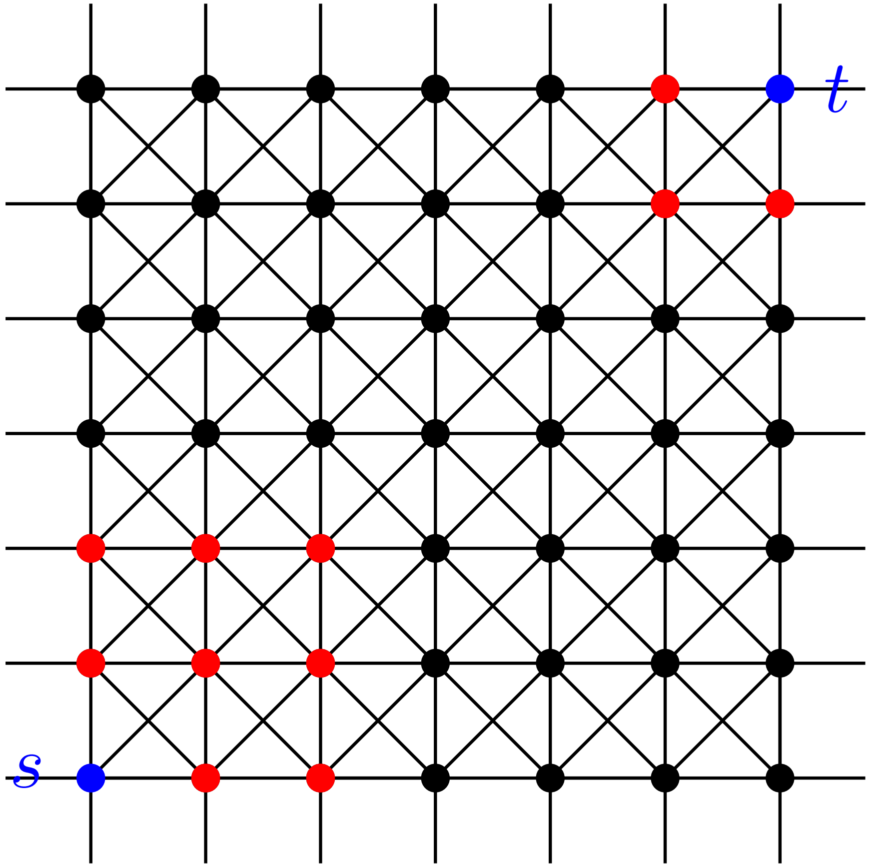
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



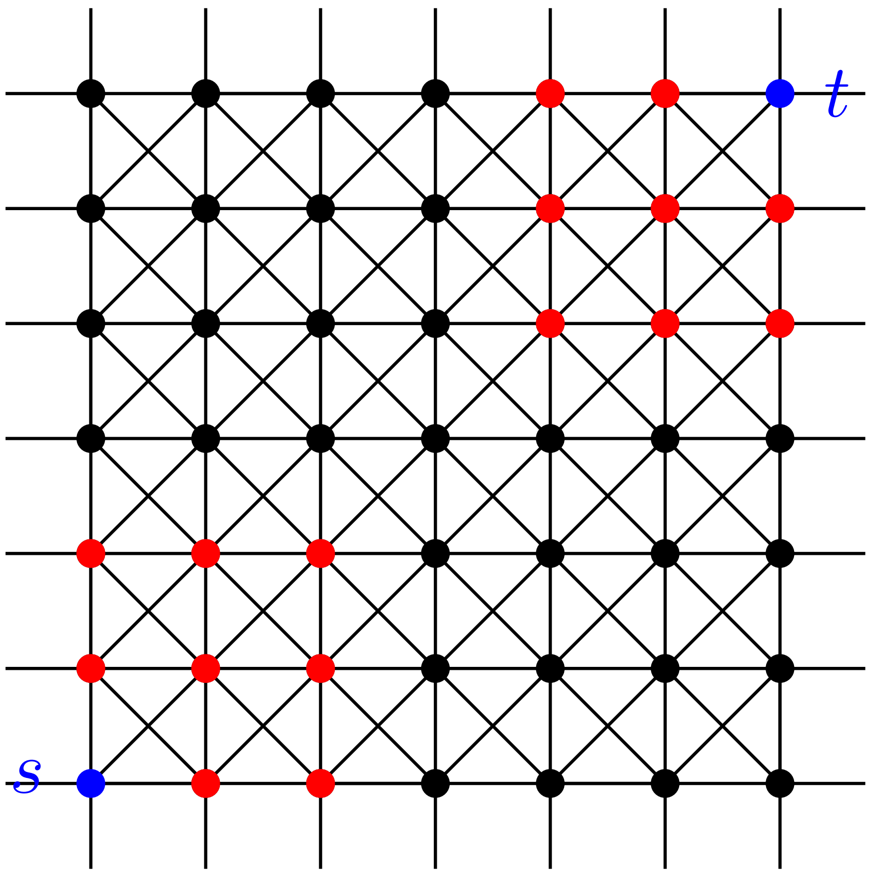
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



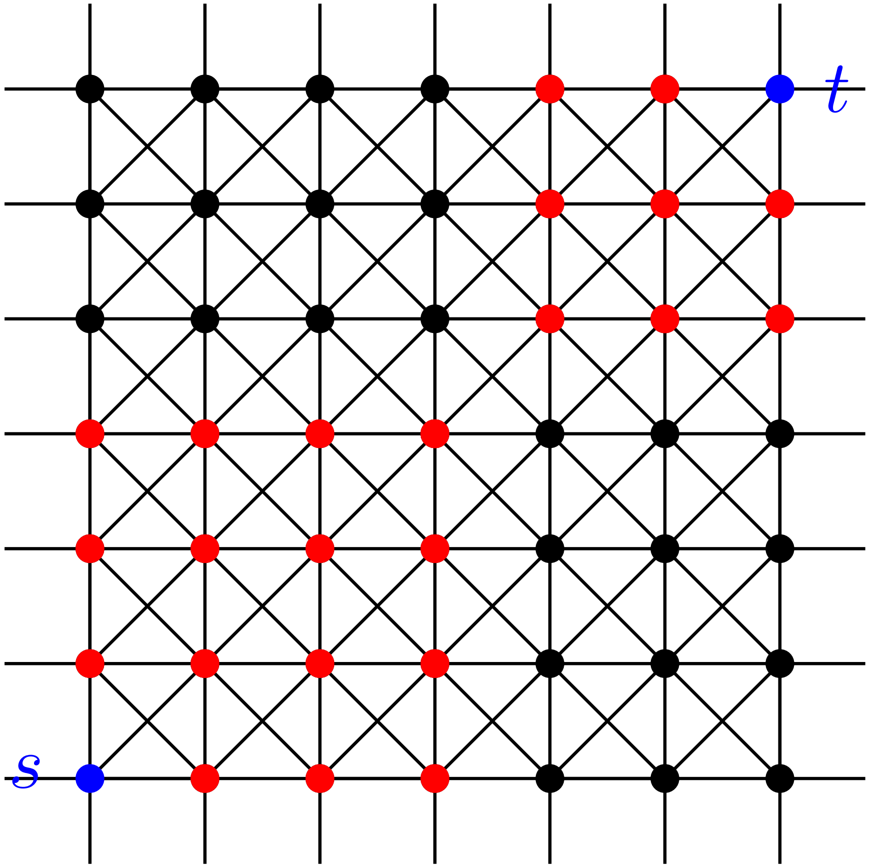
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



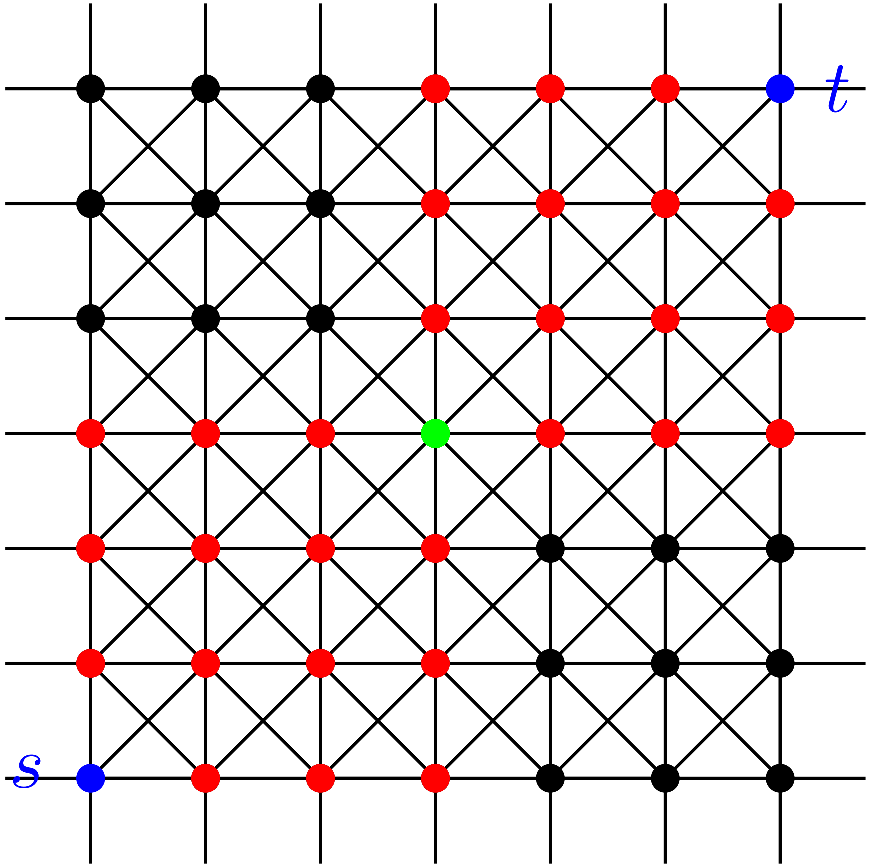
Our Idea: Balanced Bidirectional BFS

Bidirectional BFS



Our Idea: Balanced Bidirectional BFS

Bidirectional BFS

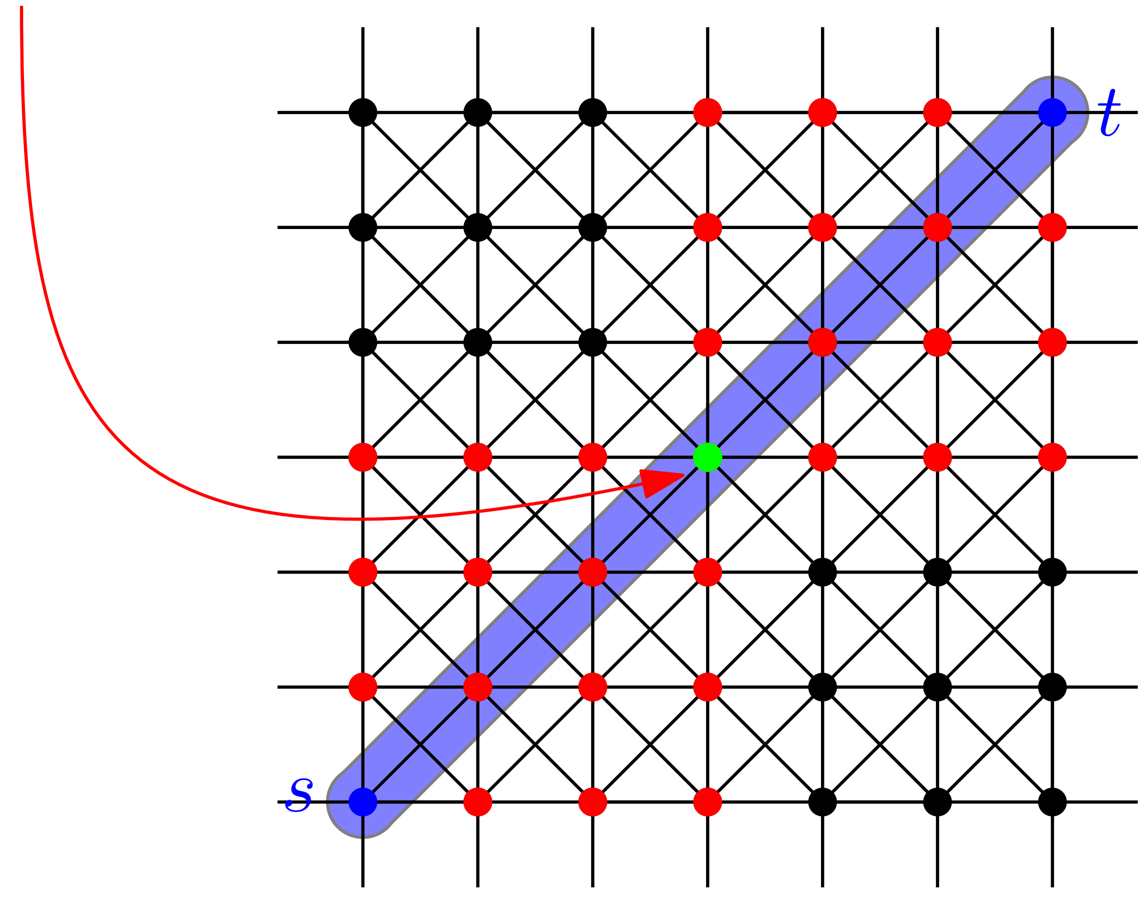


Our Idea: Balanced Bidirectional BFS

Bidirectional BFS

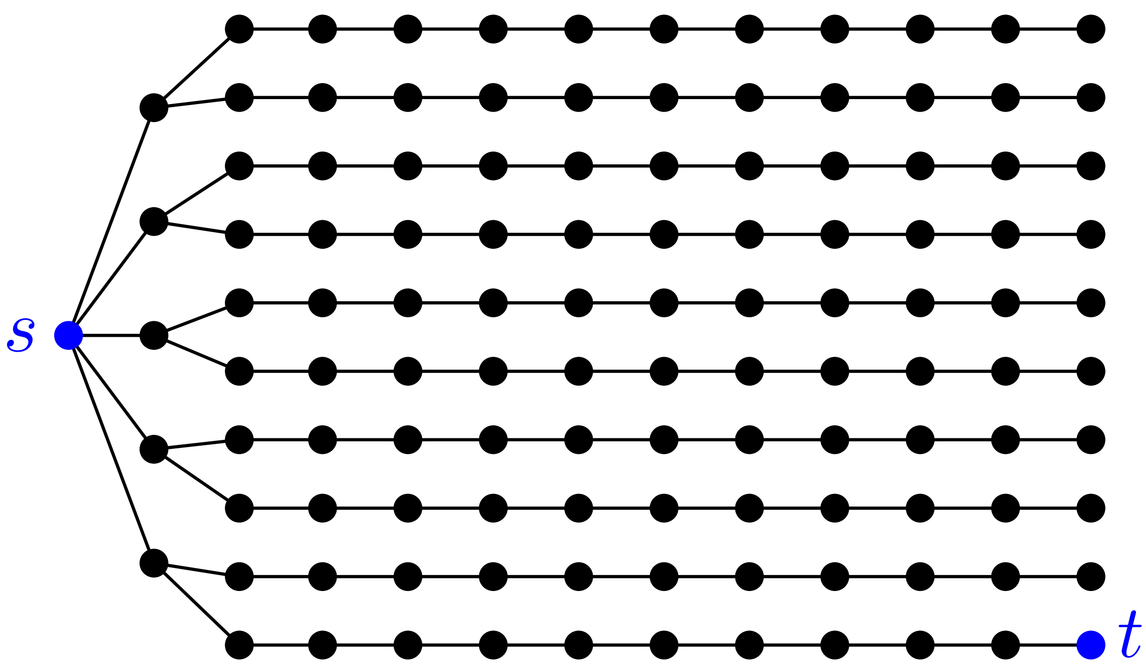
{BFS intersection} $\subseteq$ {any st-paths}

{any st-path} $\cap$ {BFS intersection} $\neq \emptyset$



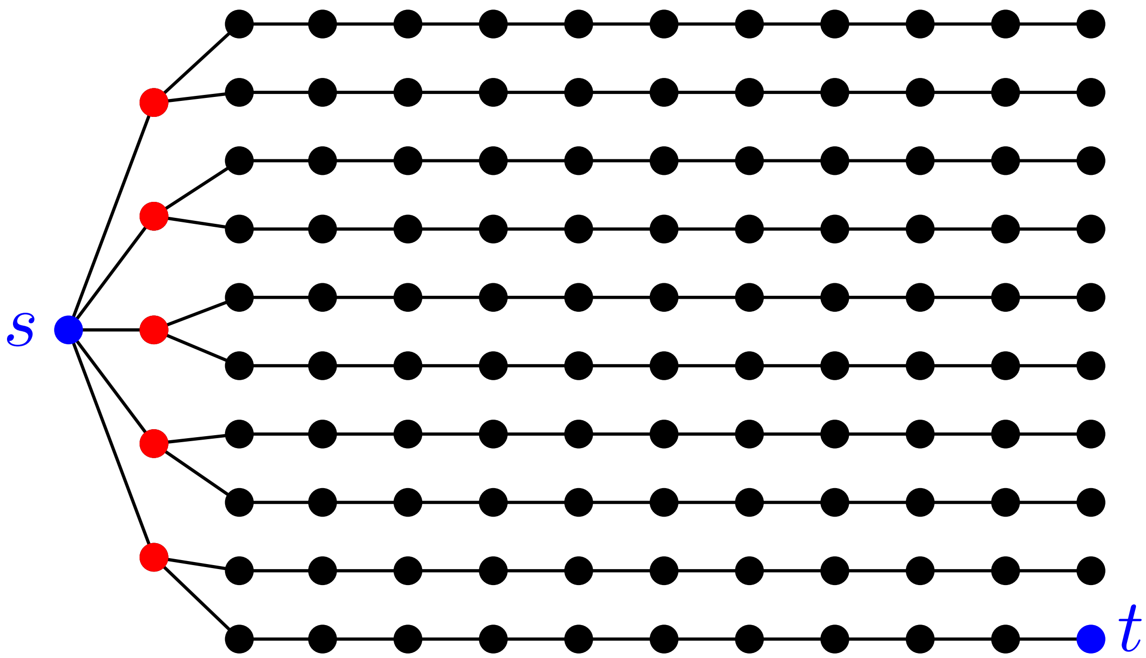
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



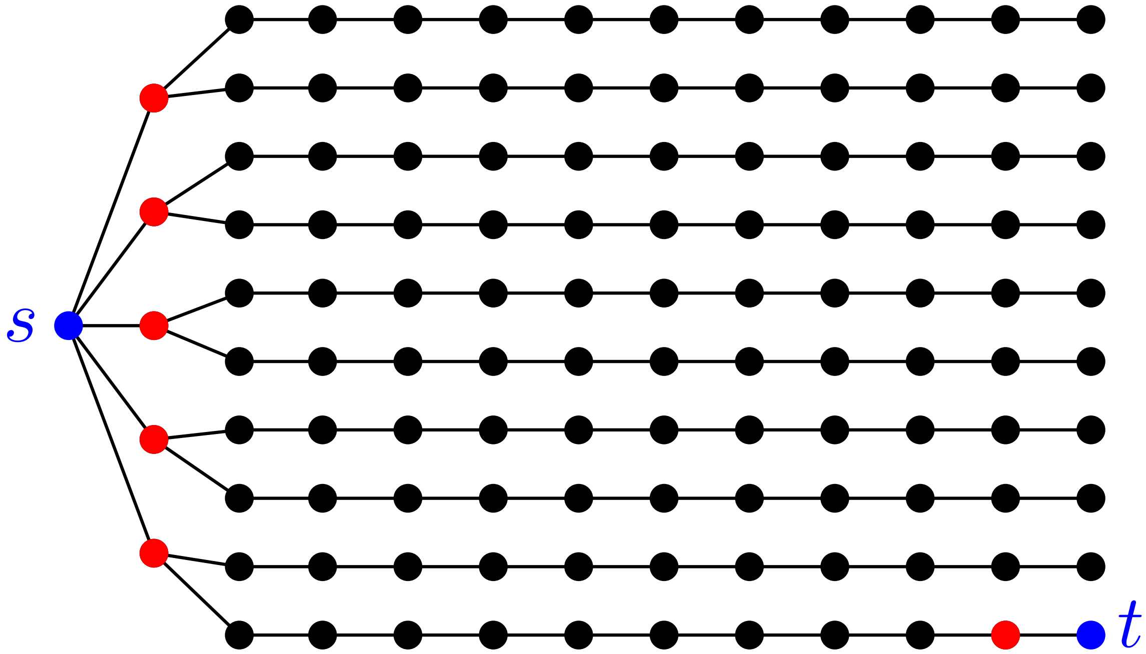
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



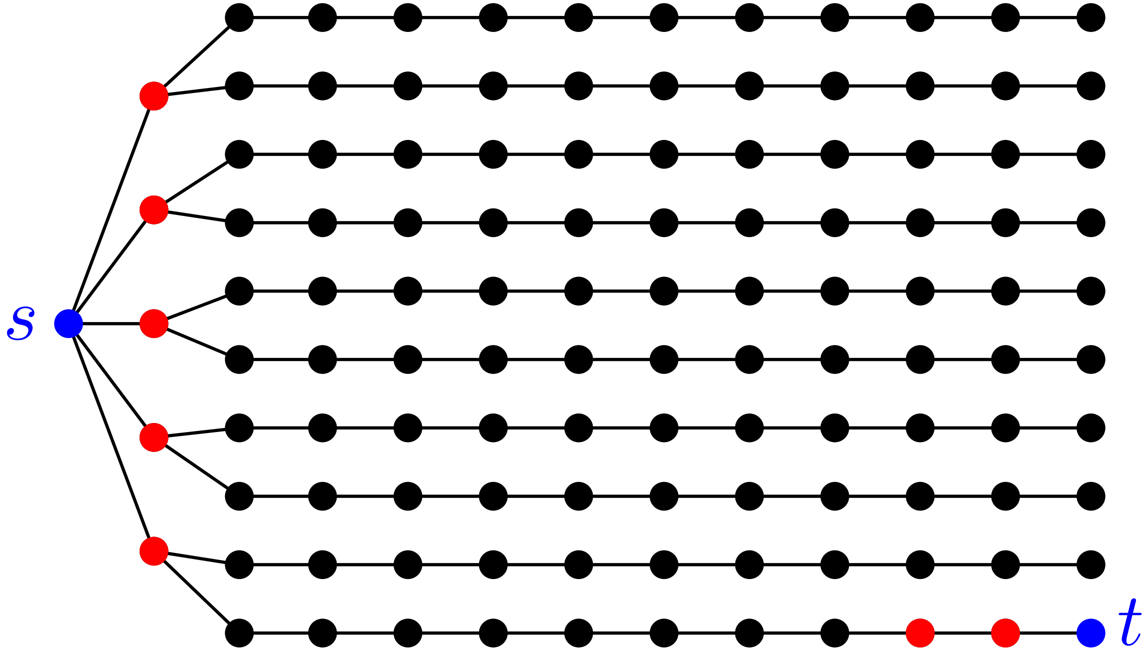
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



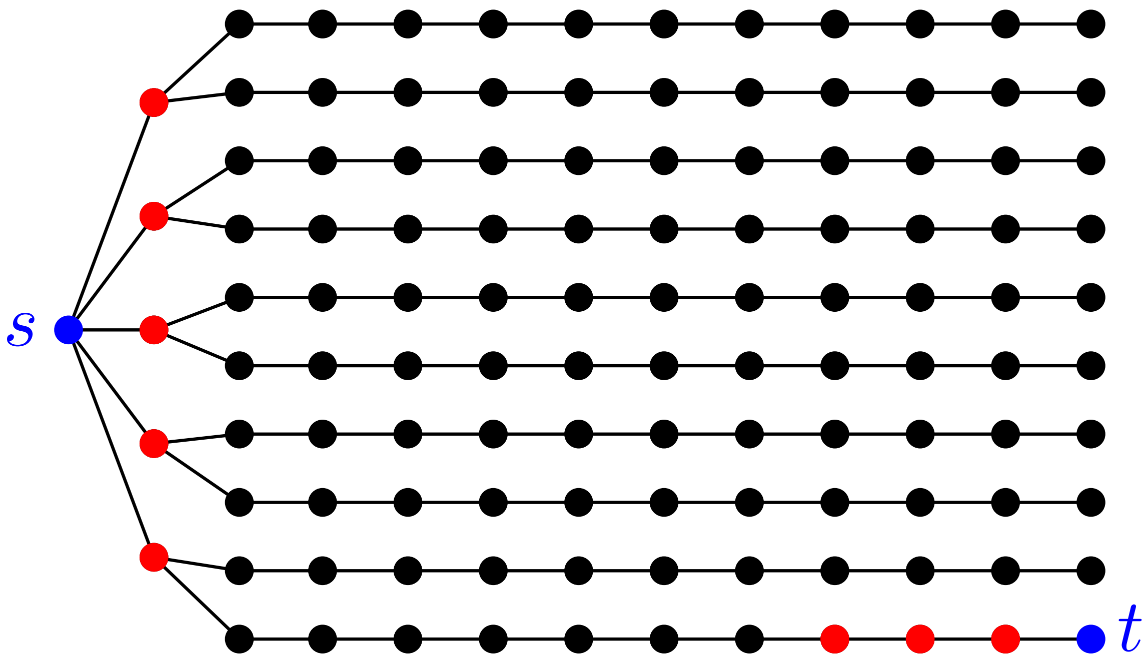
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



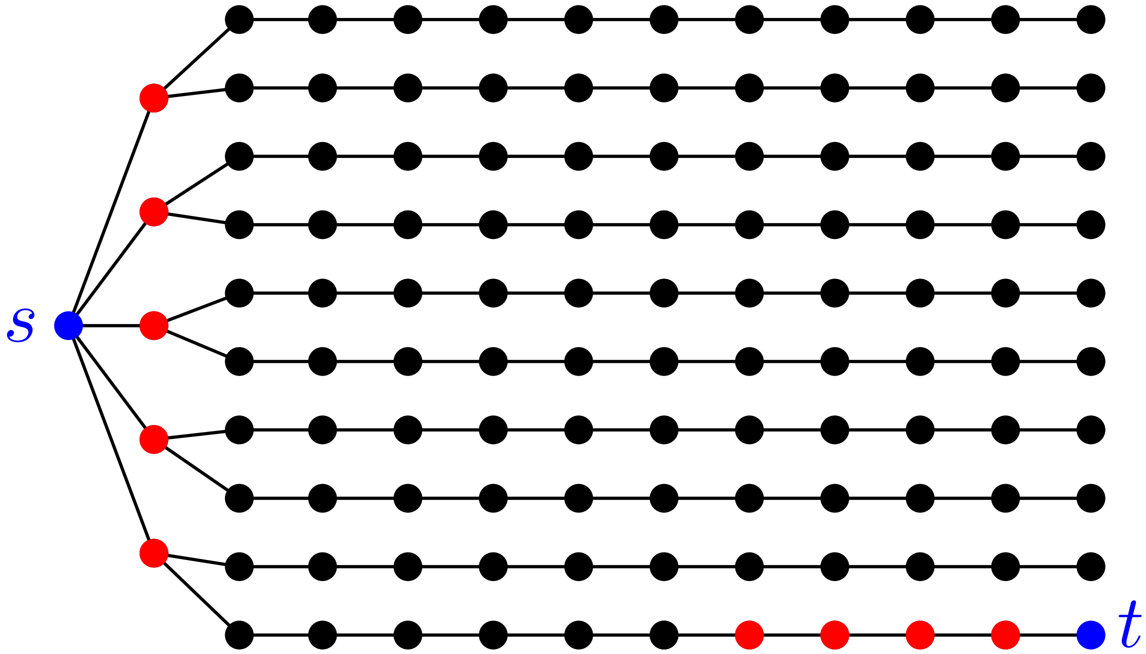
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



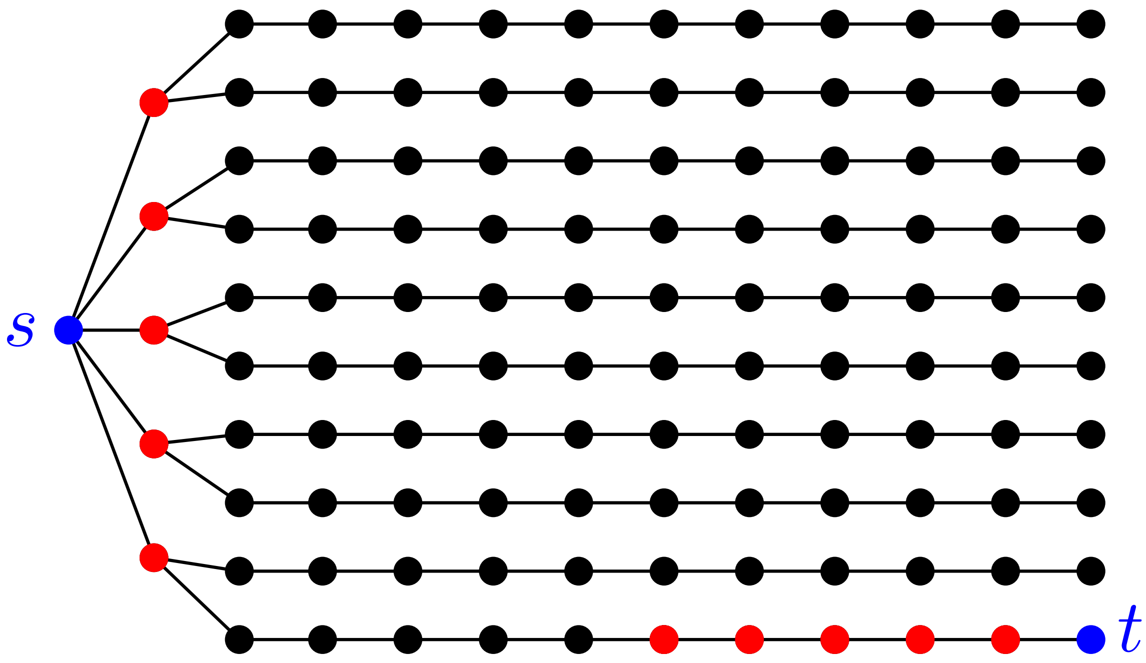
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



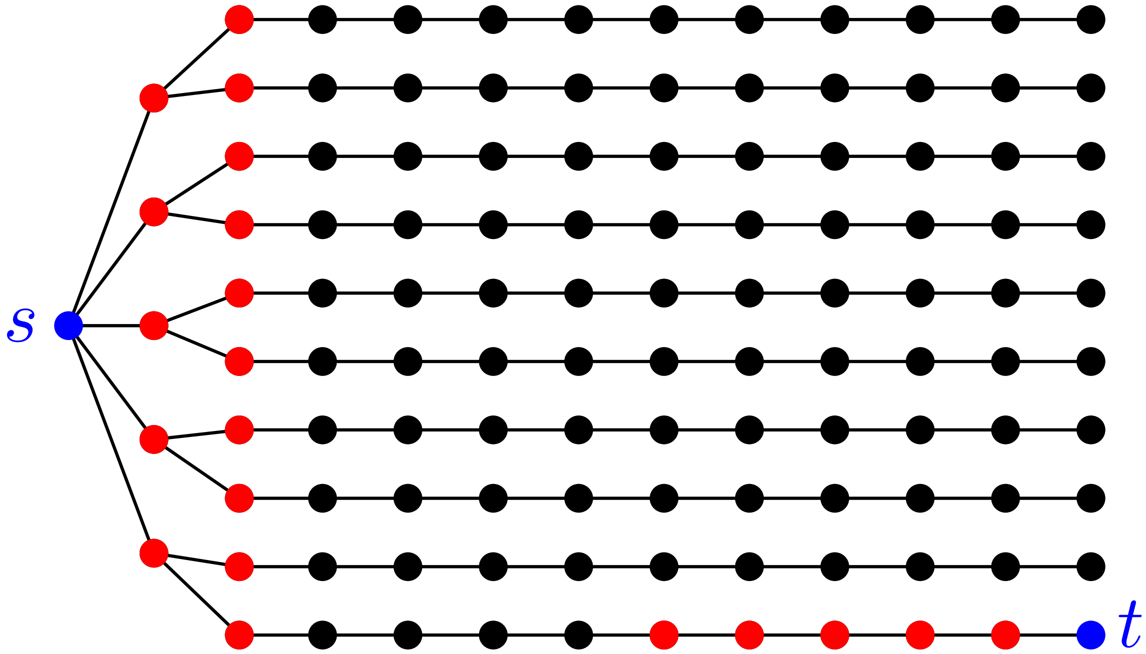
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



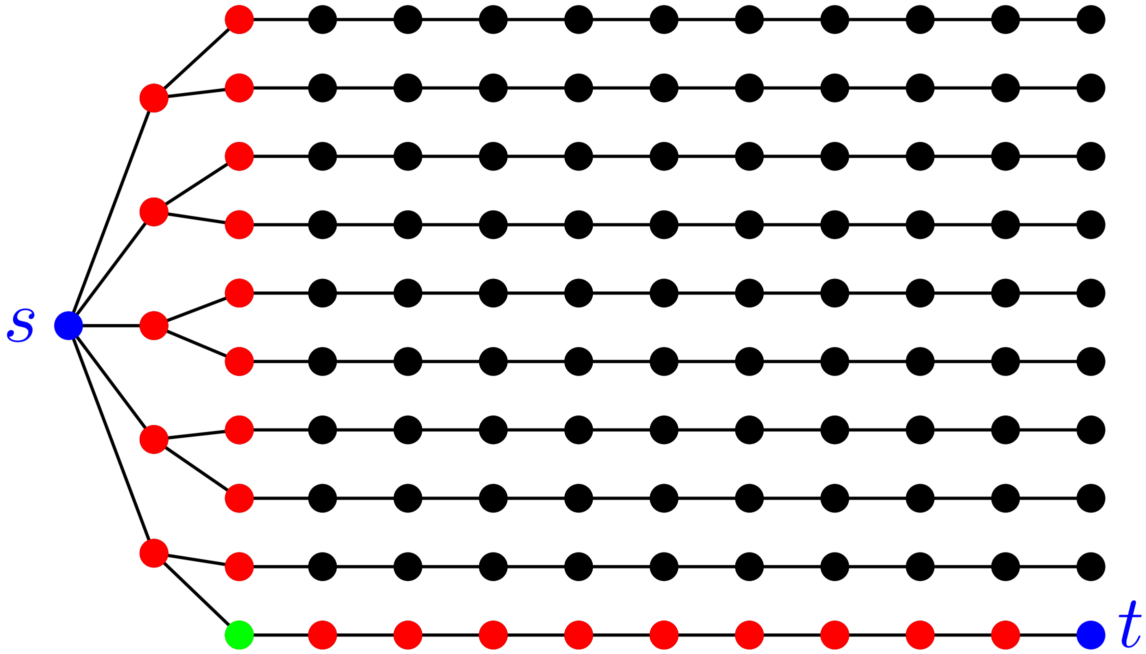
Our Idea: Balanced Bidirectional BFS

Balanced Bidirectional BFS (BBBFS)



Our Idea: Balanced Bidirectional BFS

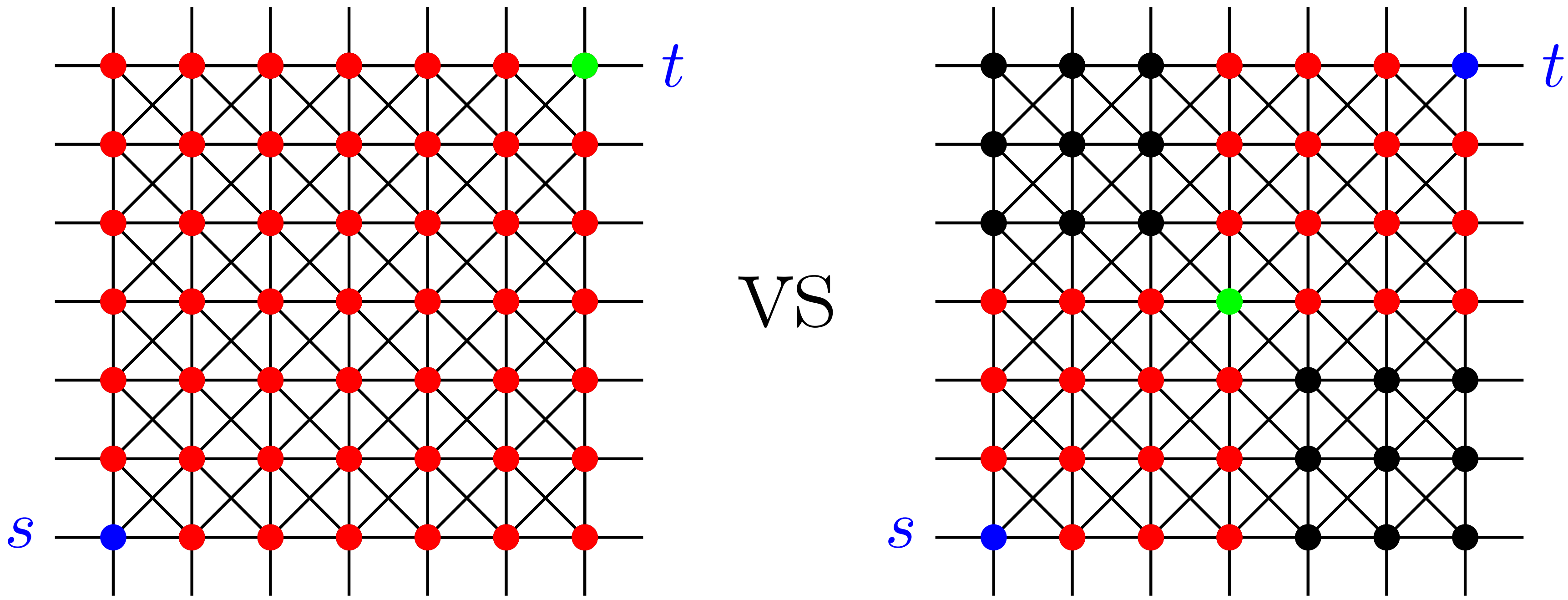
Balanced Bidirectional BFS (BBBFS)



Our Idea: Balanced Bidirectional BFS

Speed-up of BBBFS vs simple BFS?

"You may get a factor 2...
Not worth the complications!"



BBBFS on Complex Networks

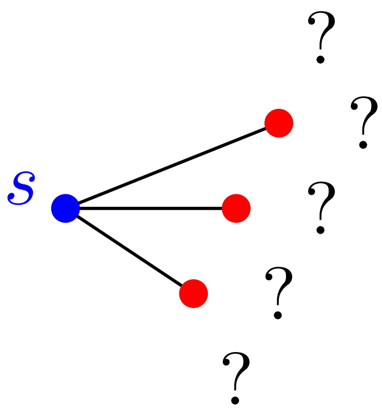
Complex Networks $\approx$ “good” Random Graph Models

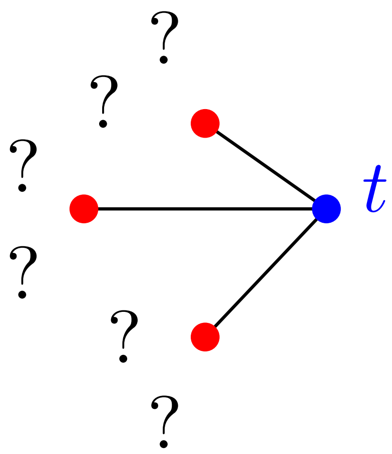
BBBFS on Complex Networks

Complex Networks $\approx$ "good" Random Graph Models

BBBFS on Complex Networks

Complex Networks $\approx$ "good" Random Graph Models



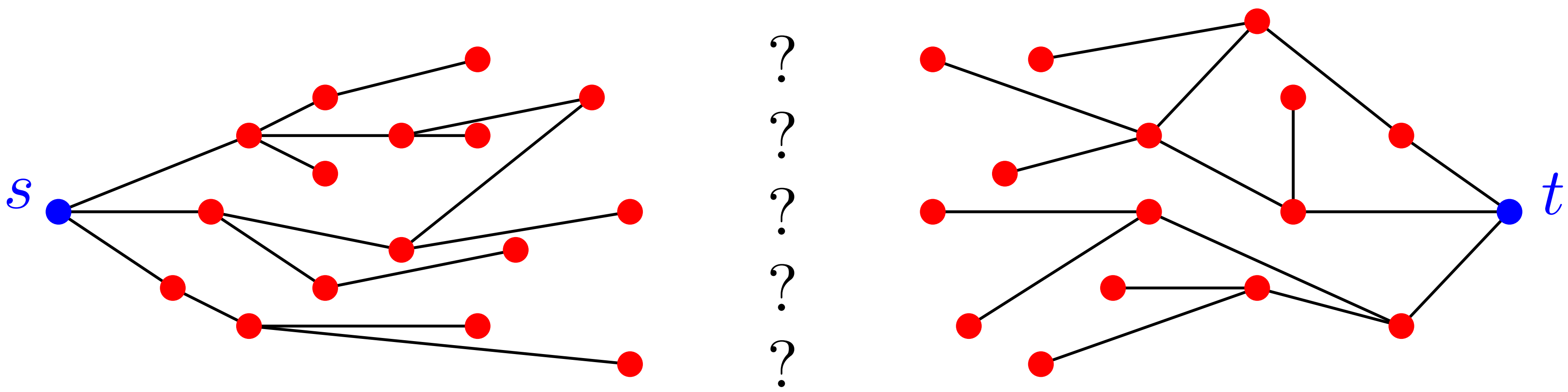


BBBFS on Complex Networks

Complex Networks $\approx$ “good” Random Graph Models

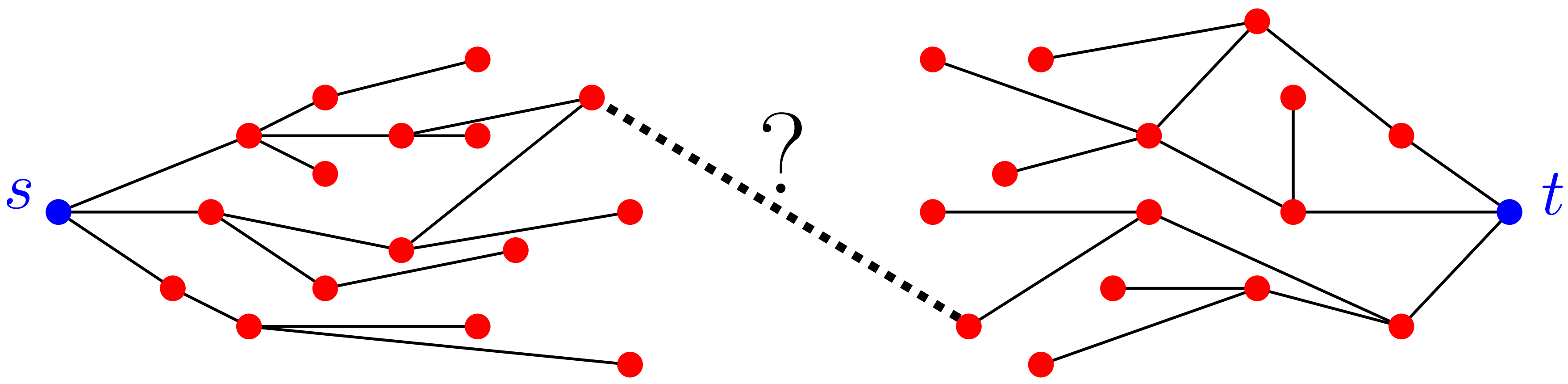
BBBFS on Complex Networks

Complex Networks $\approx$ "good" Random Graph Models



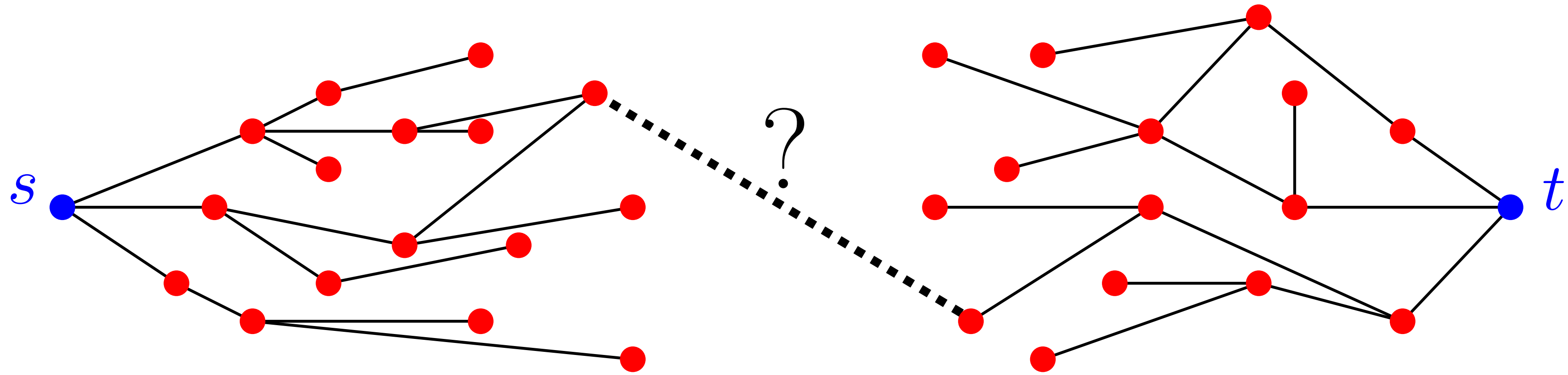
BBBFS on Complex Networks

Complex Networks $\approx$ “good” Random Graph Models



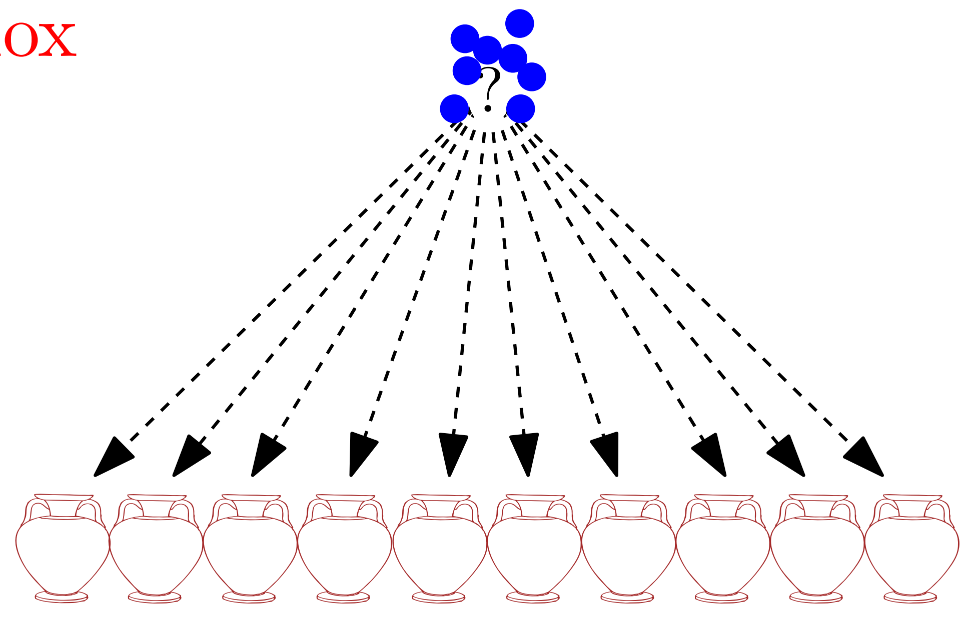
BBBFS on Complex Networks

Complex Networks $\approx$ "good" Random Graph Models



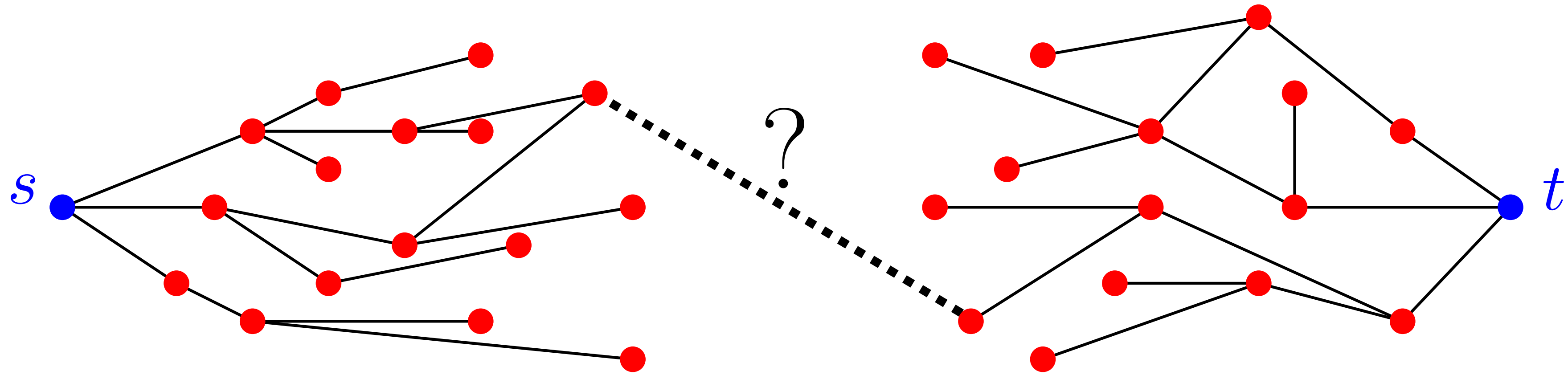
The Birthday (pseudo)Paradox

m balls u.a.r. in n bins:
Probability p of ≥ 2
balls in one bin?



BBBFS on Complex Networks

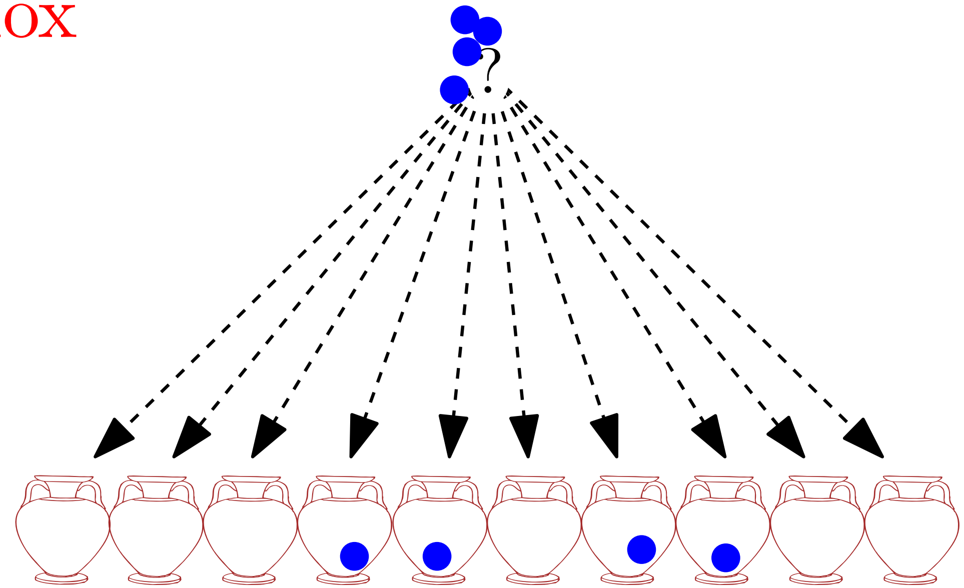
Complex Networks ≈ "good" Random Graph Models



The Birthday (pseudo)Paradox

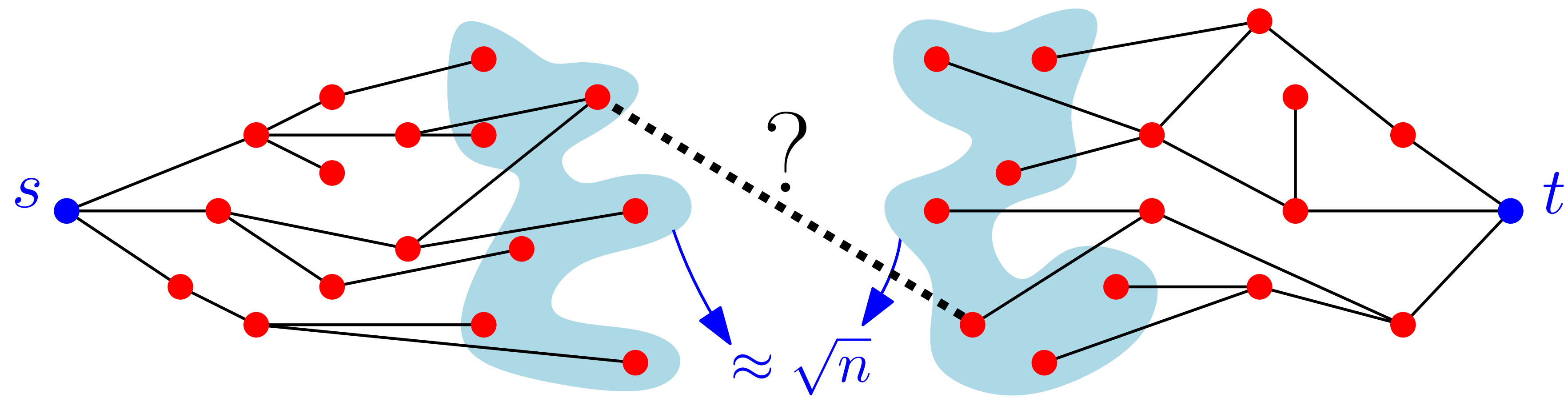
m balls u.a.r. in n bins:
Probability p of ≥ 2
balls in one bin?

$$1 - p \leq \left(1 - \frac{m}{2n}\right)^{\frac{m}{2}}$$



BBBFS on Complex Networks

Complex Networks ≈ “good” Random Graph Models

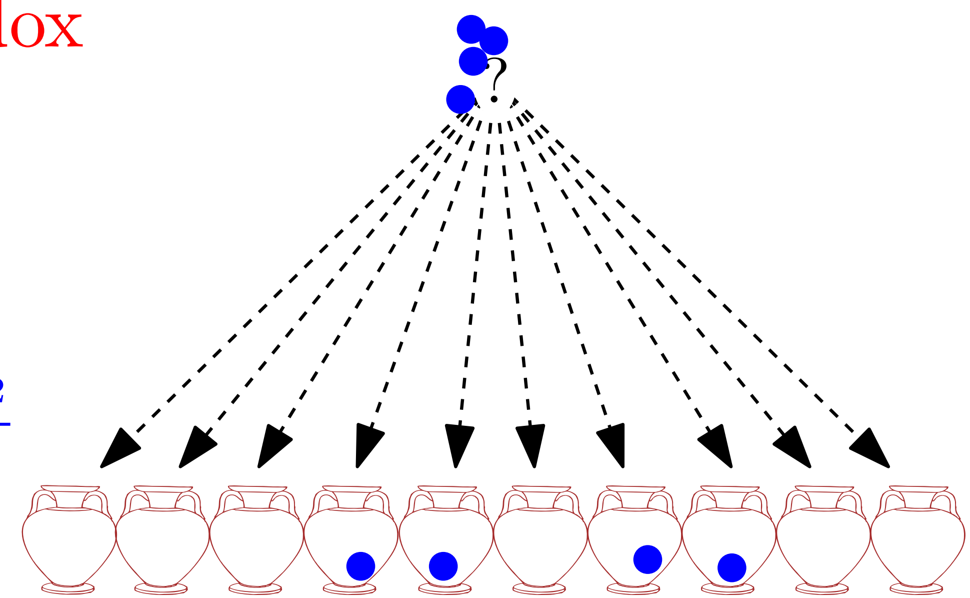


The Birthday (pseudo)Paradox

m balls u.a.r. in n bins:
Probability p of ≥ 2
balls in one bin?

$$1 - p \leq \left(1 - \frac{m}{2n}\right)^{\frac{m}{2}} \approx e^{-\frac{c^2}{4}}$$

$$m = c\sqrt{n}$$

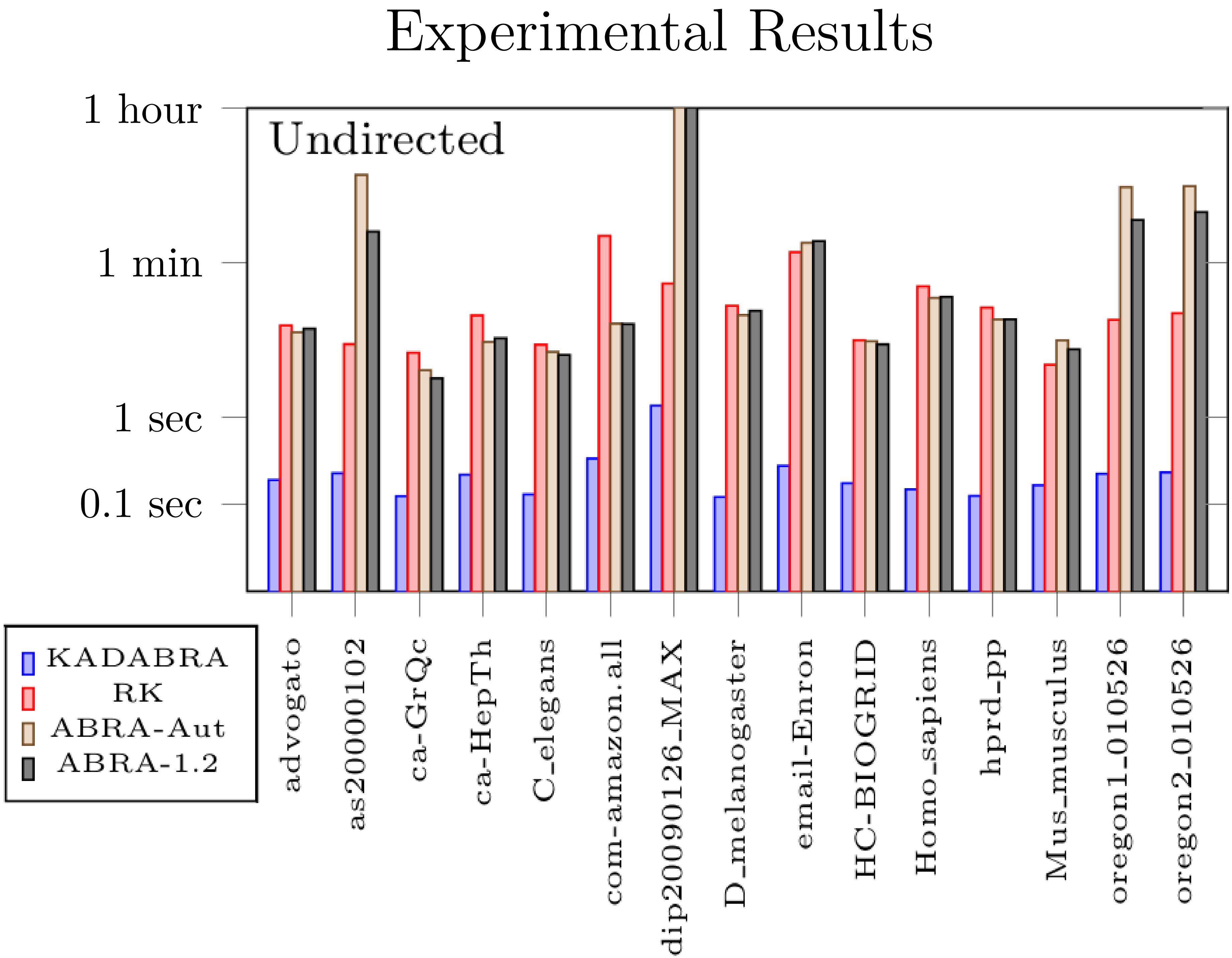


BBBFS on Random Graphs

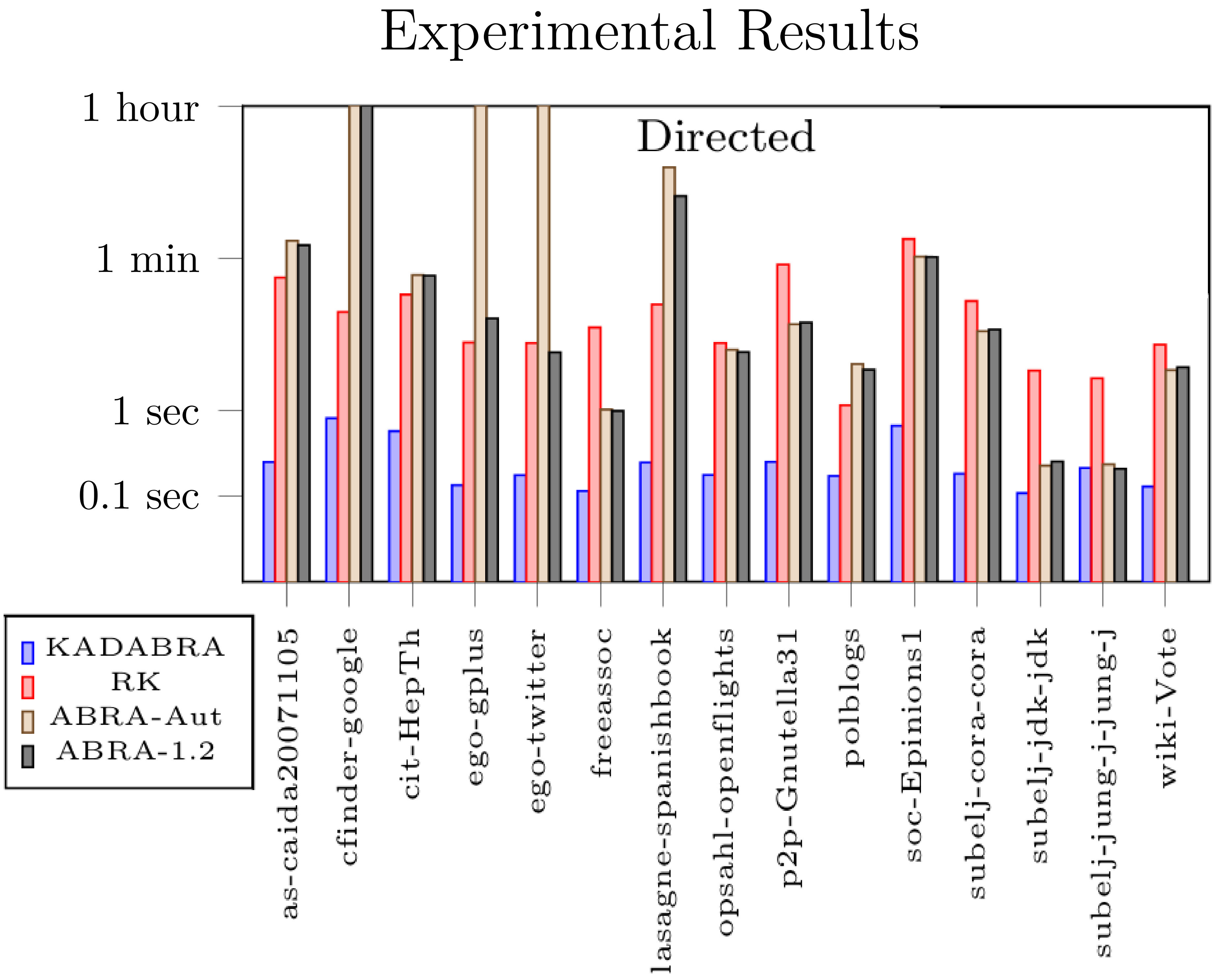
Theorem. Let G be a graph generated by one of the following models:

- the Configuration Model,
- the Norros-Reittu model,
- the Chung-Lu model, and the
- Generalized Random Graph model.

For each fixed $\epsilon > 0$, and for each pair of nodes s and t, w.h.p. the time needed to compute an st-shortest path through a BBBFS is $\mathcal{O}(n^{\frac{1}{2}+\epsilon})$ if the degree distribution λ has finite second moment, $\mathcal{O}(n^{\frac{4-\beta}{2}+\epsilon})$ if λ is a power law distribution with $2 < \beta < 3$.



Experimental Results
1 hour
1 min
1 sec
0.1 sec
Undirected
KADABRA
RK
ABRA-Aut
ABRA-1.2
advogato
as20000102
ca-GrQc
ca-HepTh
C_elegans
com-amazon.all
dip20090126_MAX
D_melanogaster
email-Enron
HC-BIOGRID
Homo_sapiens
hprd_pp
Mus_musculus
oregon1_010526
oregon2_010526



Experimental Results
Directed
1 hour
1 min
1 sec
0.1 sec
KADABRA
RK
ABRA-Aut
ABRA-1.2
as-caida20071105
cfinder-google
cit-HepTh
ego-gplus
ego-twitter
freeassoc
lasagne-spanishbook
opsahl-openflights
p2p-Gnutella31
polblogs
soc-Epinions1
subelj-cora-cora
subelj-jdk-jdk
subelj-jung-j-jung-j
wiki-Vote

Experimental Results

Wikipedia graph ($|V| = 4229697$, $|E| = 102165832$)

Rank	Page	Lower	$\tilde{\mathbf{b}}$	Upper
1)	USA	0.046278	0.047173	0.048084
2)	France	0.019522	0.020103	0.020701
3)	UK	0.017983	0.018540	0.019115
4)	England	0.016348	0.016879	0.017428
5-6)	Poland	0.012092	0.012287	0.012486
5-6)	Germany	0.011930	0.012124	0.012321
7)	India	0.009683	0.010092	0.010518
8-12)	WWII	0.008870	0.009065	0.009265
8-12)	Russia	0.008660	0.008854	0.009053
8-12)	Italy	0.008650	0.008845	0.009045
8-12)	Canada	0.008624	0.008819	0.009018
8-12)	Australia	0.008620	0.008814	0.009013

Top-k betweenness centralities with $\delta = 0.1$ and $\lambda = 0.0002$.

Experimental Results

actors → $|V|$, common movie → $|E|$

IMDB 2014 ($|V| = 1797446$, $|E| = 145760312$)

Rank	Actor	Lower	$\tilde{\mathbf{b}}$	Upper
1)	Jeremy, Ron	0.009360	0.010058	0.010808
2)	Kaufman, Lloyd	0.005936	0.006492	0.007100
3)	Hitler, Adolf	0.004368	0.004844	0.005373
4-6)	Kier, Udo	0.003250	0.003435	0.003631
4-6)	Roberts, Eric (I)	0.003178	0.003362	0.003557
4-6)	Madsen, M. (I)	0.003120	0.003305	0.003501
7-9)	Trejo, Danny	0.002652	0.002835	0.003030
7-9)	Lee, C. (I)	0.002551	0.002734	0.002931
7-12)	Estevez, Joe	0.002350	0.002534	0.002732
9-17)	Carradine, David	0.002116	0.002296	0.002492
9-17)	von Sydow, M. (I)	0.002023	0.002206	0.002405
9-17)	Keitel, Harvey (I)	0.001974	0.002154	0.002352
10-17)	Depardieu, Gèrard	0.001763	0.001943	0.002142

Top-k betweenness centralities with $\delta = 0.1$ and $\lambda = 0.0002$.

Experimental Results

since 1999

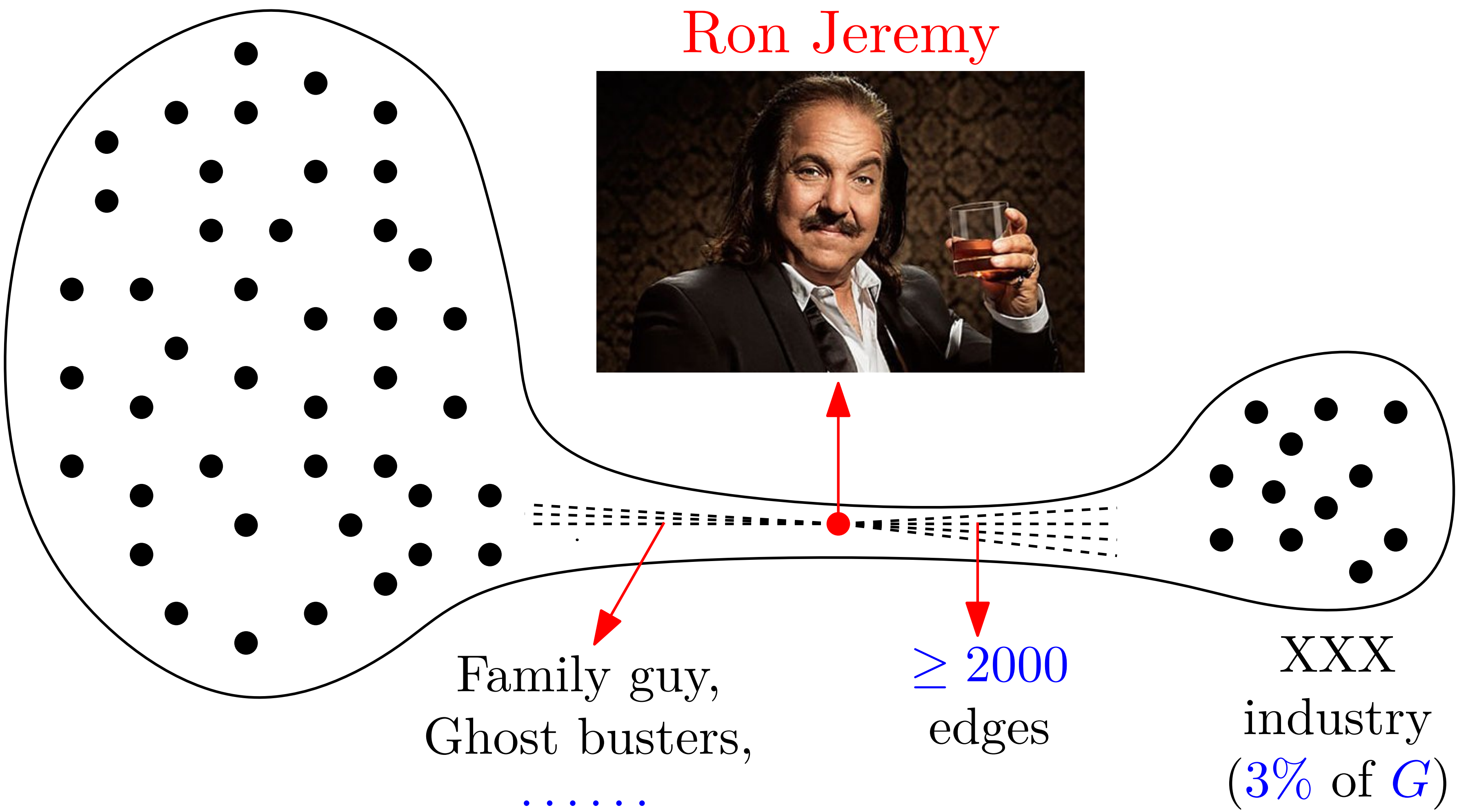
2nd from 1999 to 2009, first in 1989-94

IMDB 2014 ($|V| = 1797446$, $|E| = 145760312$)

Rank	Actor	Lower	$\tilde{\mathbf{b}}$	Upper
1)	Jeremy, Ron	0.009360	0.010058	0.010808
2)	Kaufman, Lloyd	0.005936	0.006492	0.007100
3)	Hitler, Adolf	0.004368	0.004844	0.005373
4-6)	Kier, Udo	0.003250	0.003435	0.003631
4-6)	Roberts, Eric (I)	0.003178	0.003362	0.003557
4-6)	Madsen, M. (I)	0.003120	0.003305	0.003501
7-9)	Trejo, Danny	0.002652	0.002835	0.003030
7-9)	Lee, C. (I)	0.002551	0.002734	0.002931
7-12)	Estevez, Joe	0.002350	0.002534	0.002732
9-17)	Carradine, David	0.002116	0.002296	0.002492
9-17)	von Sydow, M. (I)	0.002023	0.002206	0.002405
9-17)	Keitel, Harvey (I)	0.001974	0.002154	0.002352
10-17)	Depardieu, Gèrard	0.001763	0.001943	0.002142

Top-k betweenness centralities with $\delta = 0.1$ and $\lambda = 0.0002$.

Experimental Results



Conclusions

Take-home message(s):

- Bidirectional balanced BFS is worth trying.
- (Beware of stochastic dependence).

Conclusions

Take-home message(s):

- Bidirectional balanced BFS is worth trying.
- (Beware of stochastic dependence).

Open Problem: Improve the algorithm (combine ABRA and KADABRA) and...

Conclusions

Take-home message(s):

- Bidirectional balanced BFS is worth trying.
- (Beware of stochastic dependence).

Open Problem: Improve the algorithm (combine ABRA and KADABRA) and...

...

Conclusions

Take-home message(s):

- Bidirectional balanced BFS is worth trying.
- (Beware of stochastic dependence).

Open Problem: Improve the algorithm (combine ABRA and KADABRA) and...

Find an acronim for *ALAKAZAM*!

Thank You!