Four degrees of separation in 69 billion friendships

Paolo Boldi   Marco Rosa **Sebastiano Vigna**
Laboratory for Web Algorithmics
Università degli Studi di Milano, Italy
Lars Backstrom, Johan Ugander
Facebook

First intuition

Literature!

First intuition
Literature!

- Frigyes Karinthy, in his 1929 short story “Láncszemek” (“Chains'”) suggested that any two persons are distanced by at most six friendship links

First intuition

Literature!

- Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains'") suggested that any two persons are distanced by at most six friendship links
- Just an (optimistic) positivistic statement about combinatorial explosion

First intuition
Literature!

- Frigyes Karinthy, in his 1929 short story “Láncszemek” (“Chains'”) suggested that any two persons are distanced by at most six friendship links
- Just an (optimistic) positivistic statement about combinatorial explosion
- Used by John Guare's in his 1990 eponymous play (and movie by Fred Shepisi)

The Sociologists

The Sociologists

- M. Kochen, I. de Sola Pool: *Contacts and influences.* (Manuscript, early 50s)

The Sociologists

- M. Kochen, I. de Sola Pool: *Contacts and influences.* (Manuscript, early 50s)
- A. Rapoport, W.J. Horvath: *A study of a large sociogram.* (Behav.Sci. 1961)

The Sociologists

- M. Kochen, I. de Sola Pool: *Contacts and influences.* (Manuscript, early 50s)
- A. Rapoport, W.J. Horvath: *A study of a large sociogram.* (Behav.Sci. 1961)
- S. Milgram, *An experimental study of the small world problem.* (Sociometry, 1969)

Milgram’s question

Milgram's question

- What is the distance distribution of the acquaintance graph?

Milgram's question

- What is the distance distribution of the acquaintance graph?
- That is, how many pairs of people are friends, how many are not friends but have a friend in common, etc

Milgram's question

- What is the distance distribution of the acquaintance graph?
- That is, how many pairs of people are friends, how many are not friends but have a friend in common, etc
- Note: sociologists measure the *degrees of separation* (i.e., the number of intermediaries); computer scientists measure the graph-theoretic distance (just add one)

Milgram’s experiment

Milgram’s experiment

- ~300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)

Milgram's experiment

- ~300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker

Milgram's experiment

- ~300 people (*starting population*) are asked to dispatch a parcel to a single individual (*target*)
- The target was a Boston stockbroker
- The starting population is selected as follows:
  - ~100 were random Boston inhabitants (group A)
  - ~100 were random Nebraska stockbrokers (group B)
  - ~100 were random Nebraska inhabitants (group C)

Milgram’s experiment

Milgram’s experiment

- Rules of the game:

Milgram's experiment

- Rules of the game:
  - parcels could be directly sent *only* to someone the sender knows personally ("first-name acquaintance")

Milgram’s experiment

- Rules of the game:
  - parcels could be directly sent *only* to someone the sender knows personally (“first-name acquaintance”)
  - 453 intermediaries happened to be involved in the experiments (besides the starting population and the target)

Milgram’s experiment

Milgram's experiment

- Actually completed: 22%

Milgram's experiment

- Actually completed: 22%
- Average distance *of the completed chains* in the range 5.4 to 6.7 (depending on the group)

Milgram's experiment

- Actually completed: 22%
- Average distance *of the completed chains* in the range 5.4 to 6.7 (depending on the group)
- 6.7 (i.e., 5.7 degrees of separation) was the average distance of the random group

How difficult is it…

How difficult is it...

- ...to reproduce (at least the easy part of) Milgram's experiment on a large scale?

How difficult is it...

- ...to reproduce (at least the easy part of) Milgram's experiment on a large scale?
- i.e.: how can one compute or approximate the distance distribution of a given *huge* graph?

How difficult is it...

- ...to reproduce (at least the easy part of) Milgram's experiment on a large scale?
- i.e.: how can one compute or approximate the distance distribution of a given *huge* graph?
- (given, of course, that one has a *huge* friendship graph...)

Graph distances and distribution

Graph distances and distribution

- Given a graph, $d(x,y)$ is the length of the shortest path from x to y (∞ if one cannot go from x to y)

Graph distances and distribution

- Given a graph, $d(x,y)$ is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, $d(x,y)=d(y,x)$

Graph distances and distribution

- Given a graph, $d(x,y)$ is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, $d(x,y)=d(y,x)$
- For every t, count the number of pairs (x,y) such that $d(x,y)=t$

Graph distances and distribution

- Given a graph, $d(x,y)$ is the length of the shortest path from x to y (∞ if one cannot go from x to y)
- For *undirected* graphs, $d(x,y)=d(y,x)$
- For every t, count the number of pairs (x,y) such that $d(x,y)=t$
- The fraction of pairs at distance t is (the density function of) a distribution

Previous experiments:
Online Social Networks

Previous experiments: Online Social Networks

- Leskovec and Horvitz (2008) find 6.6 degrees of separation on a one-month MSN Messenger communication graph with 180 M nodes and 1.3 G edges

Previous experiments: Online Social Networks

- Leskovec and Horvitz (2008) find 6.6 degrees of separation on a one-month MSN Messenger communication graph with 180 M nodes and 1.3 G edges

- Degrees of separation in Twitter in 2010 were 3.67 on 5 G follows (but the figure is quite meaningless when links are created without permission at both ends)

Previous experiments: Online Social Networks

- Leskovec and Horvitz (2008) find 6.6 degrees of separation on a one-month MSN Messenger communication graph with 180 M nodes and 1.3 G edges
- Degrees of separation in Twitter in 2010 were 3.67 on 5 G follows (but the figure is quite meaningless when links are created without permission at both ends)
- Our largest dataset: 712 M people, 69 G friendship links

HyperANF

HyperANF

- A diffusion-based approximated algorithm that computes the distance distribution (2011)

HyperANF

- A diffusion-based approximated algorithm that computes the distance distribution (2011)
- Following ANF [Palmer *et al.*, 2002]

HyperANF

- A diffusion-based approximated algorithm that computes the distance distribution (2011)
- Following ANF [Palmer *et al.*, 2002]
- It uses HyperLogLog counters [Flajolet *et al.*, 2007] and broadword programming for low-level parallelization

Intermediate step

Intermediate step

- The neighbourhood function: for each *t*, the *number* of pairs at distance *at most t*

Intermediate step

- The neighbourhood function: for each t, the *number* of pairs at distance *at most t*
- Easy to derive the cumulative distribution function of distances (just divide by the last value)

Intermediate step

- The neighbourhood function: for each *t*, the *number* of pairs at distance *at most t*
- Easy to derive the cumulative distribution function of distances (just divide by the last value)
- Easy to derive the number of reachable pairs and probability mass function (but relative error becomes absolute error!)

How do you compute it?

How do you compute it?

- Many many breadth-first visits: $O(mn)$, needs direct access

How do you compute it?

- Many many breadth-first visits: $O(mn)$, needs direct access
- Sampling: a fraction of breadth-first visits, very unreliable results on graphs that are not strongly connected, needs direct access

How do you compute it?

- Many many breadth-first visits: $O(mn)$, needs direct access
- Sampling: a fraction of breadth-first visits, very unreliable results on graphs that are not strongly connected, needs direct access
- Edith Cohen's [JCSS 1997] size estimation framework: very powerful but does not scale or parallelize really well, needs direct access

Alternative: Diffusion

Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD ’02

Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD '02
- Let $B_t(x)$ be the ball of radius t around x (nodes at distance at most t from x)

Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD '02
- Let $B_t(x)$ be the ball of radius t around x (nodes at distance at most t from x)
- Clearly $B_0(x)=\{x\}$

Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD '02
- Let $B_t(x)$ be the ball of radius t around x (nodes at distance at most t from x)
- Clearly $B_0(x)=\{x\}$
- But also $B_{t+1}(x)=\bigcup_{x\to y}B_t(y)\cup\{x\}$

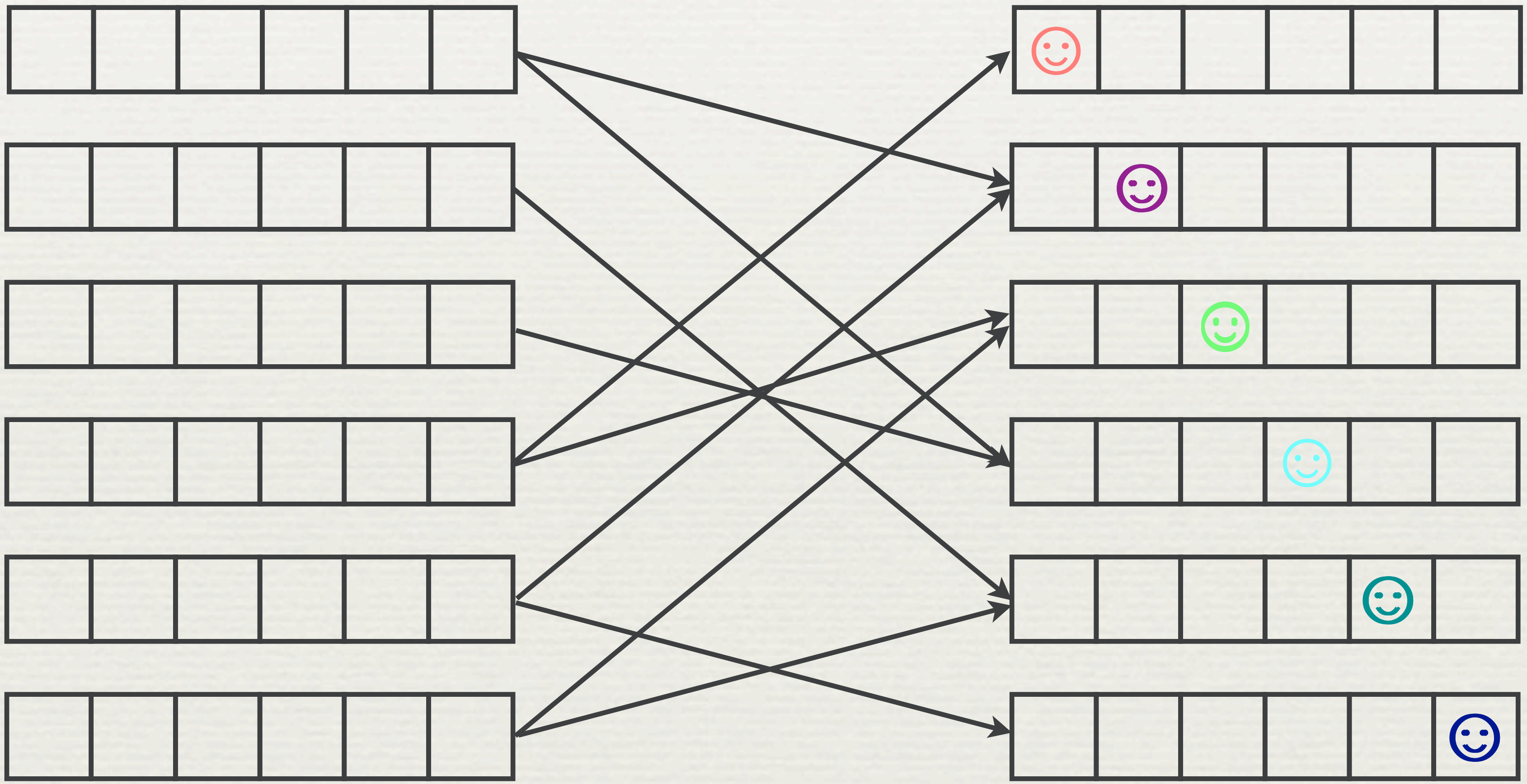
Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD '02
- Let $B_t(x)$ be the ball of radius t around x (nodes at distance at most t from x)
- Clearly $B_0(x)=\{x\}$
- But also $B_{t+1}(x)=\bigcup_{x\to y}B_t(y)\cup\{x\}$
- So we can compute balls by enumerating the arcs $x\to y$ and performing set unions

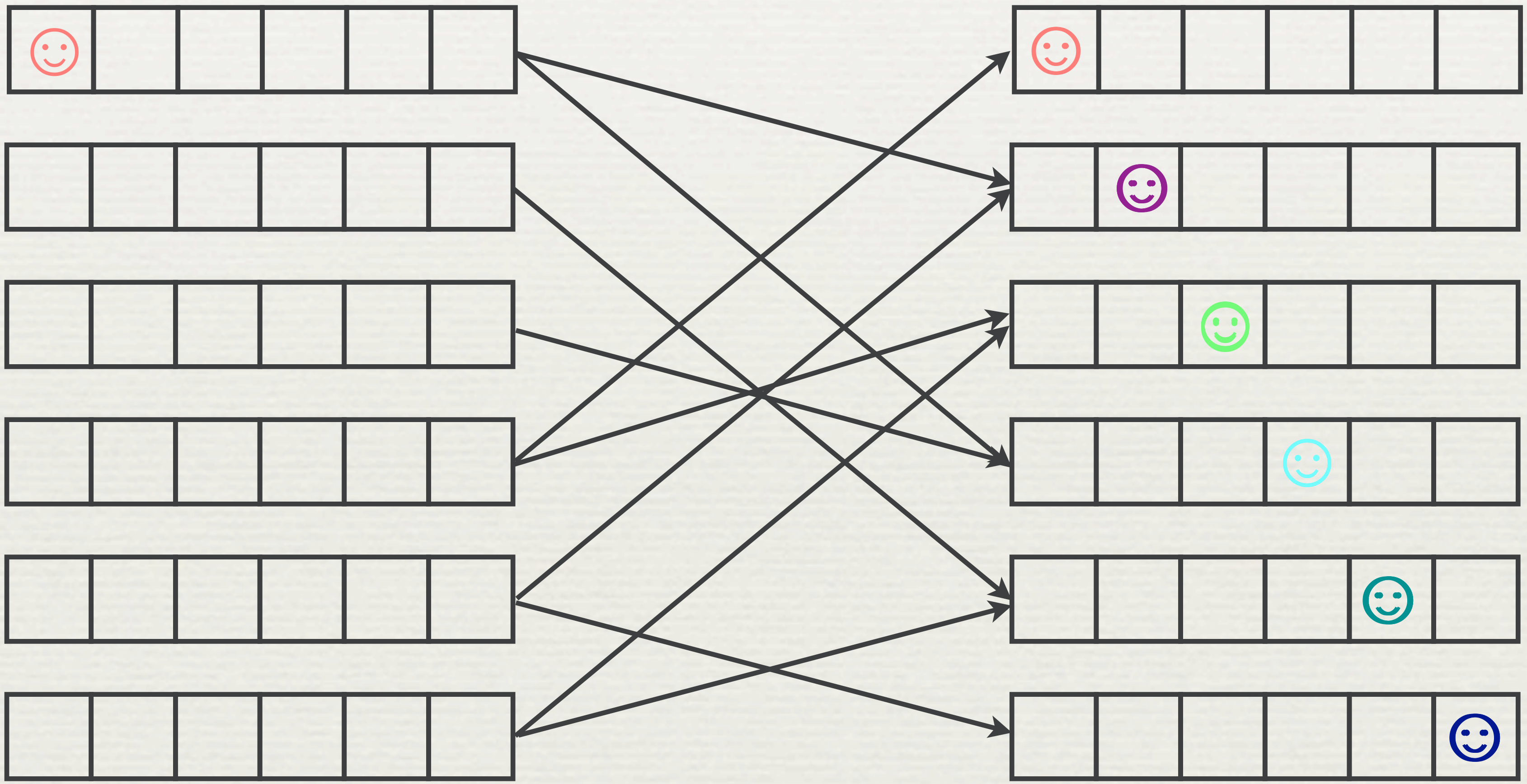
Alternative: Diffusion

- Basic idea: Palmer *et. al*, KDD '02
- Let $B_t(x)$ be the ball of radius t around x (nodes at distance at most t from x)
- Clearly $B_0(x)=\{x\}$
- But also $B_{t+1}(x)=\bigcup_{x\to y}B_t(y)\cup\{x\}$
- So we can compute balls by enumerating the arcs $x\to y$ and performing set unions
- The neighbourhood function at t is given by the sum of the sizes of the balls of radius t!

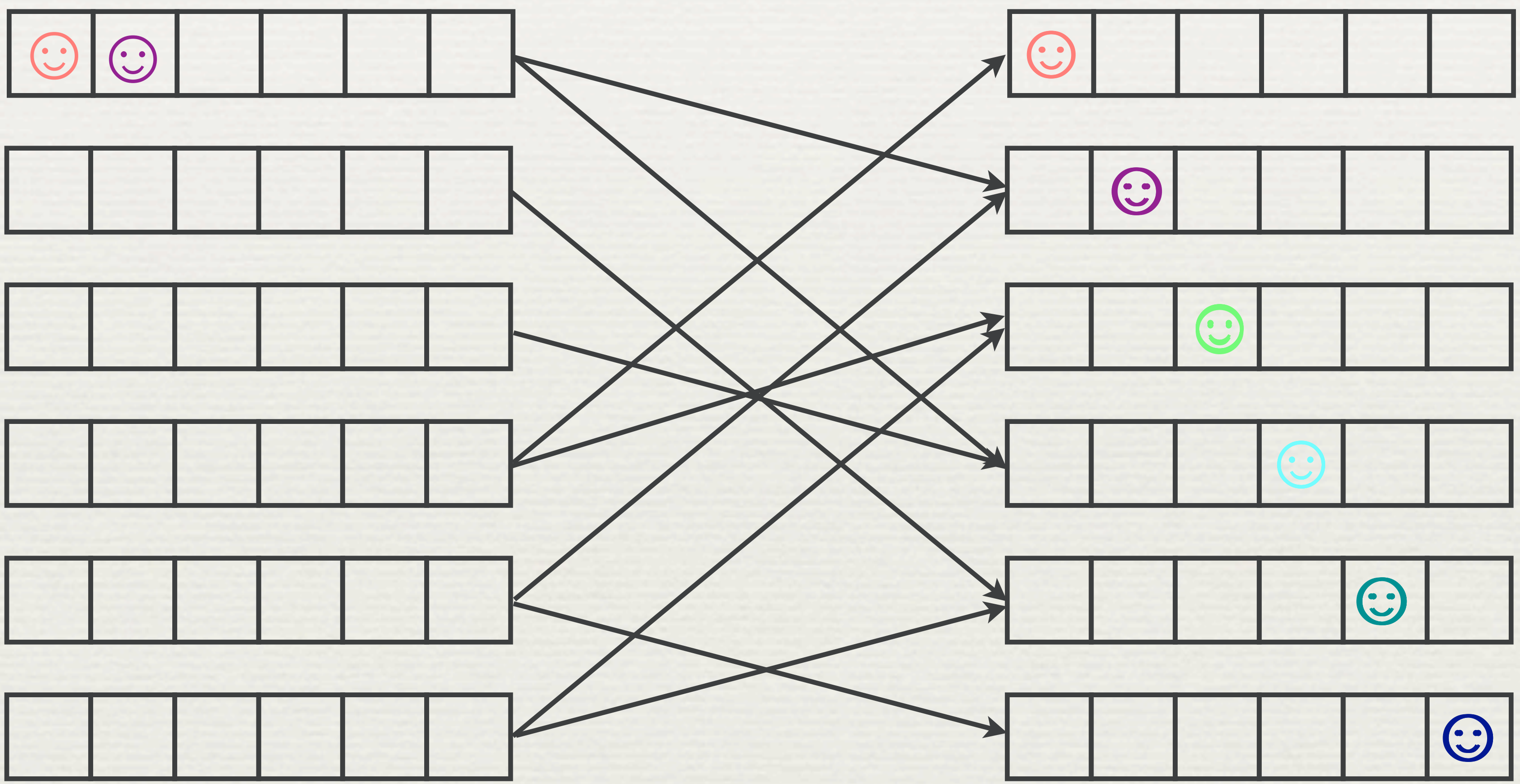
A round of updates



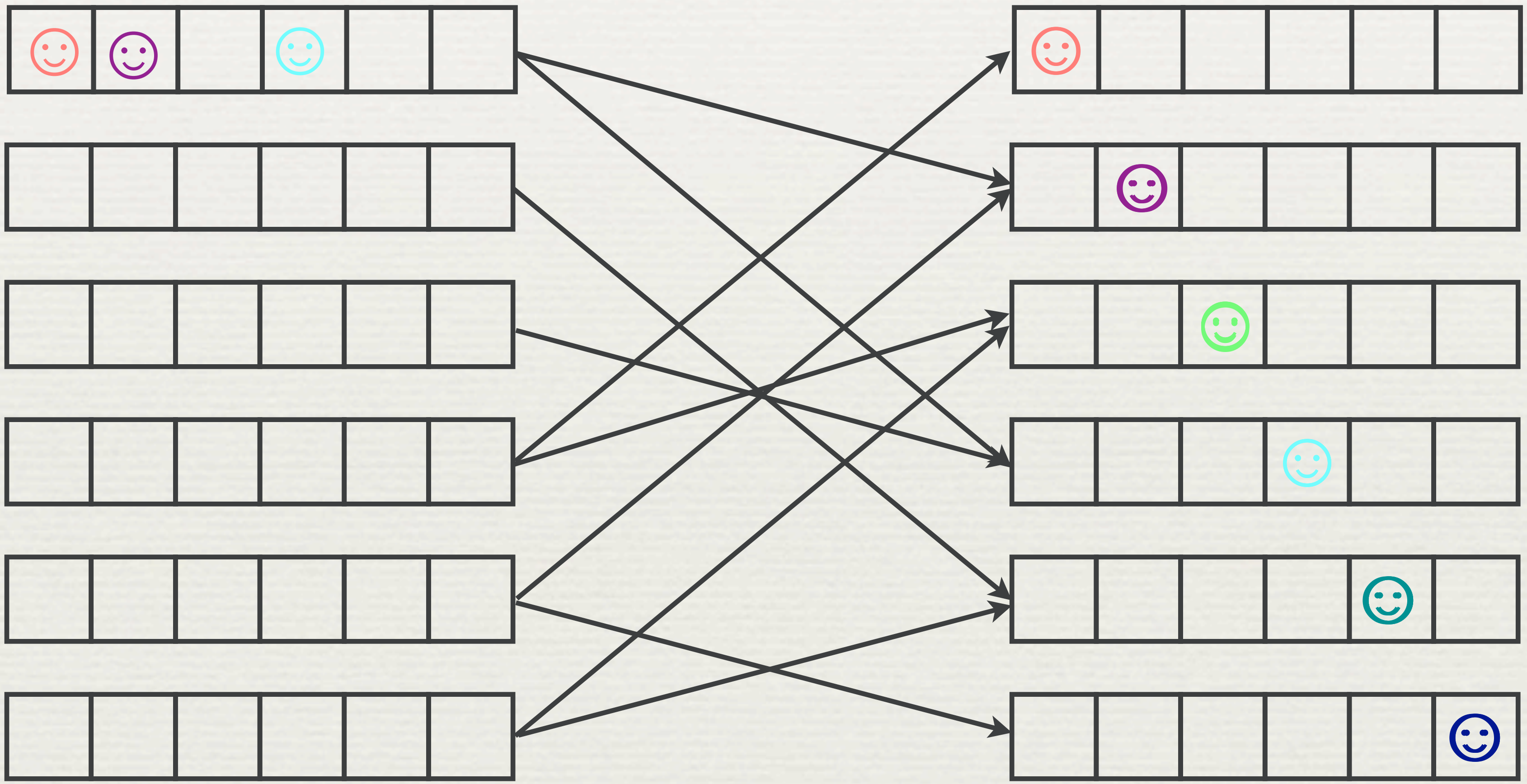
A round of updates



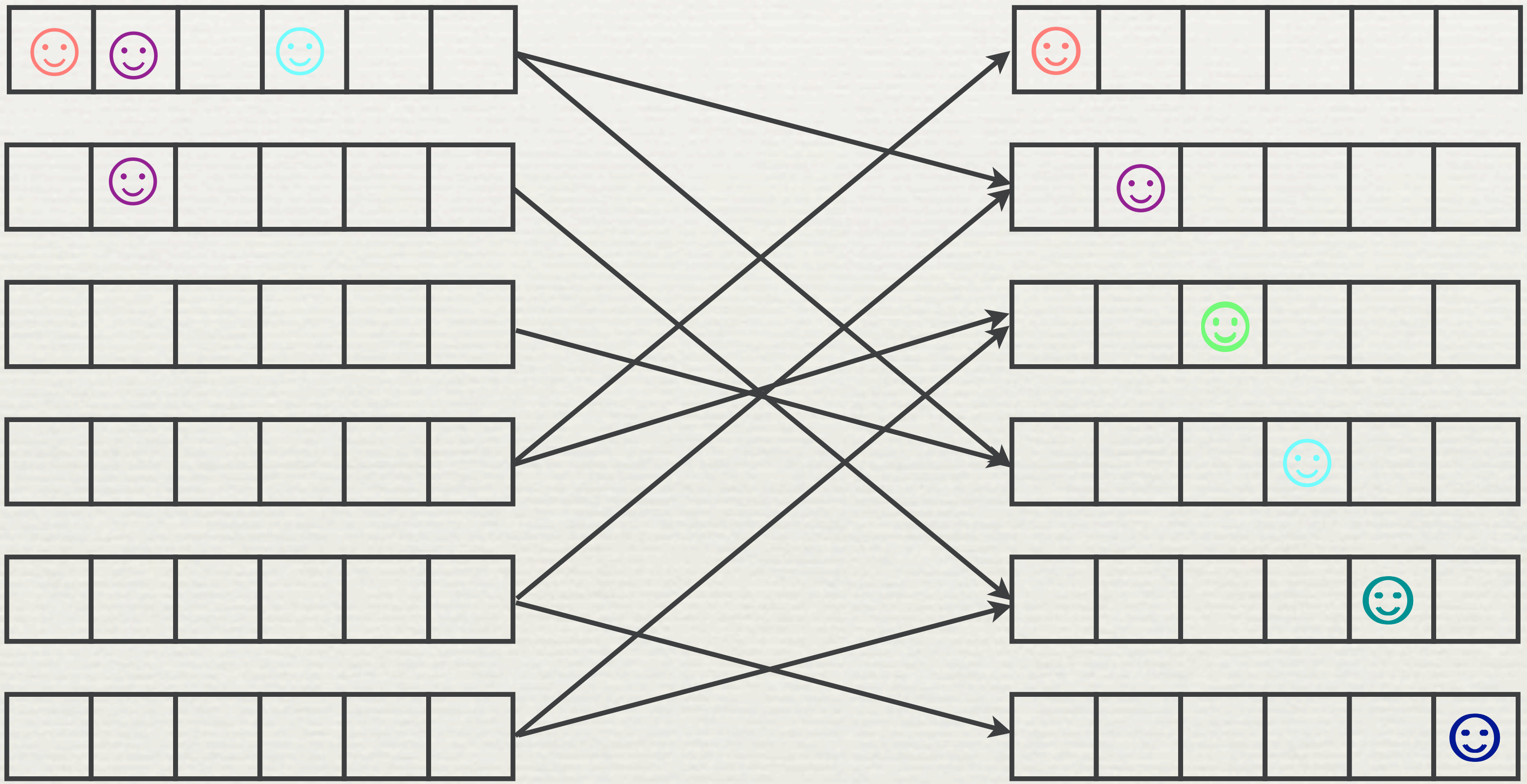
A round of updates



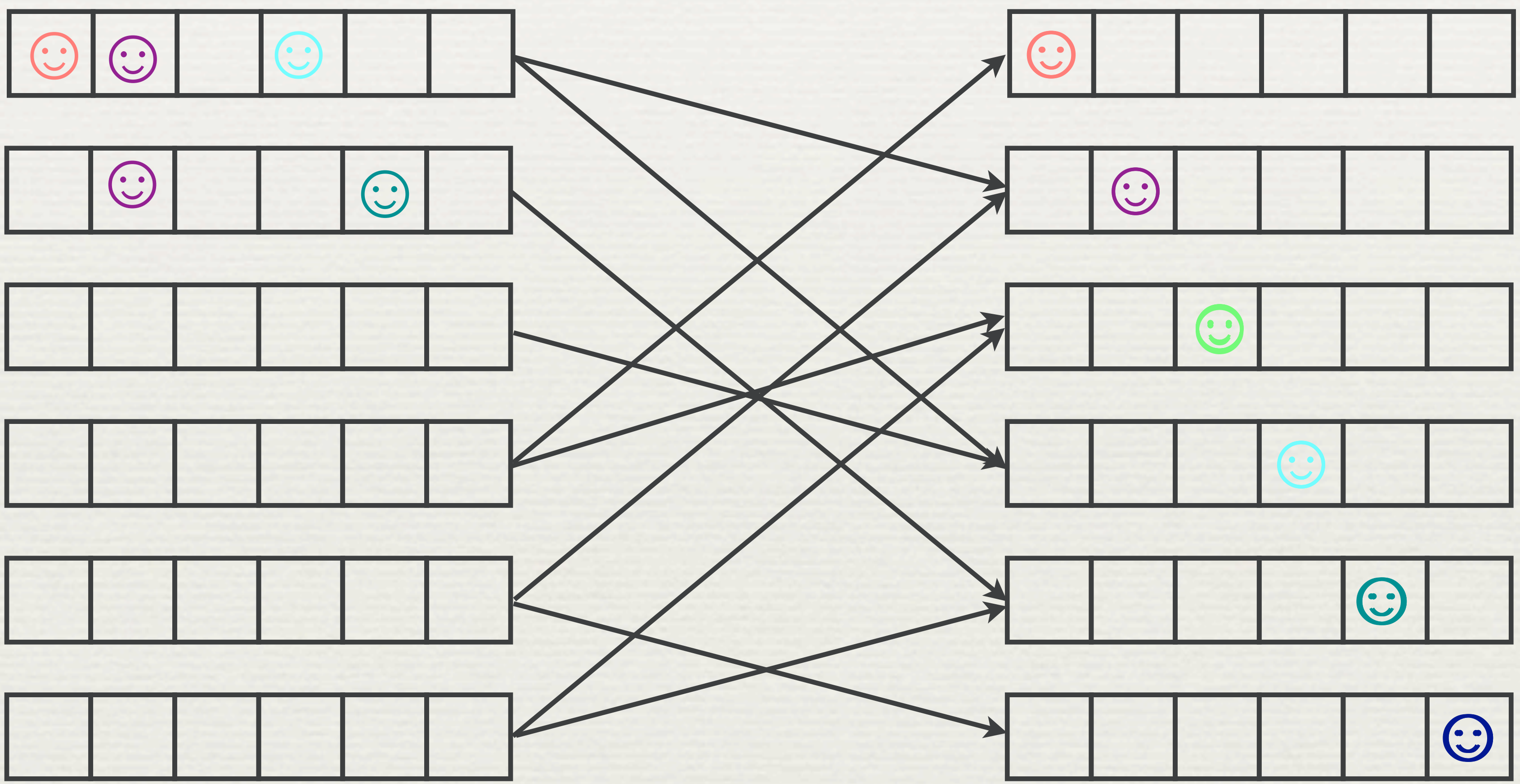
A round of updates



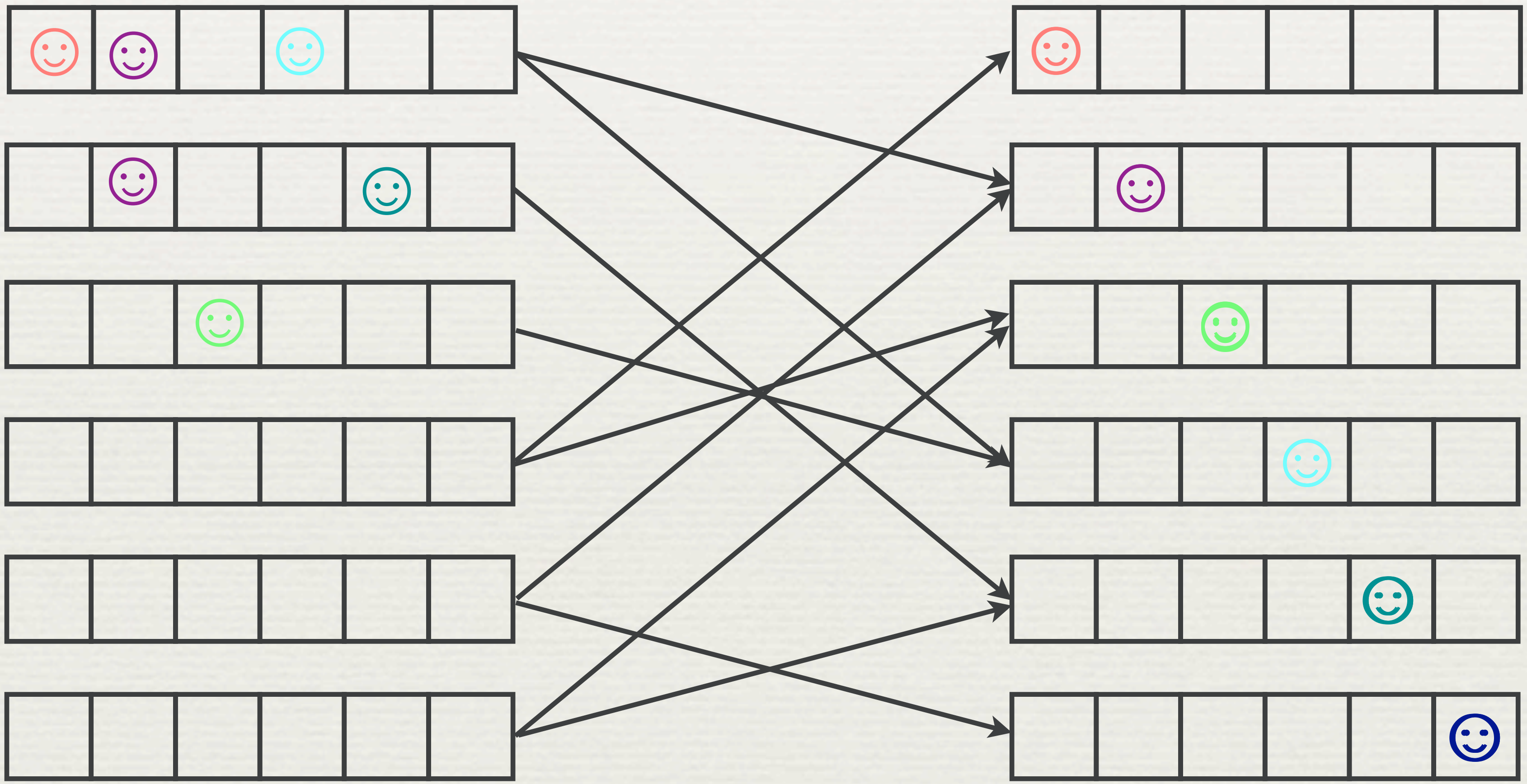
A round of updates



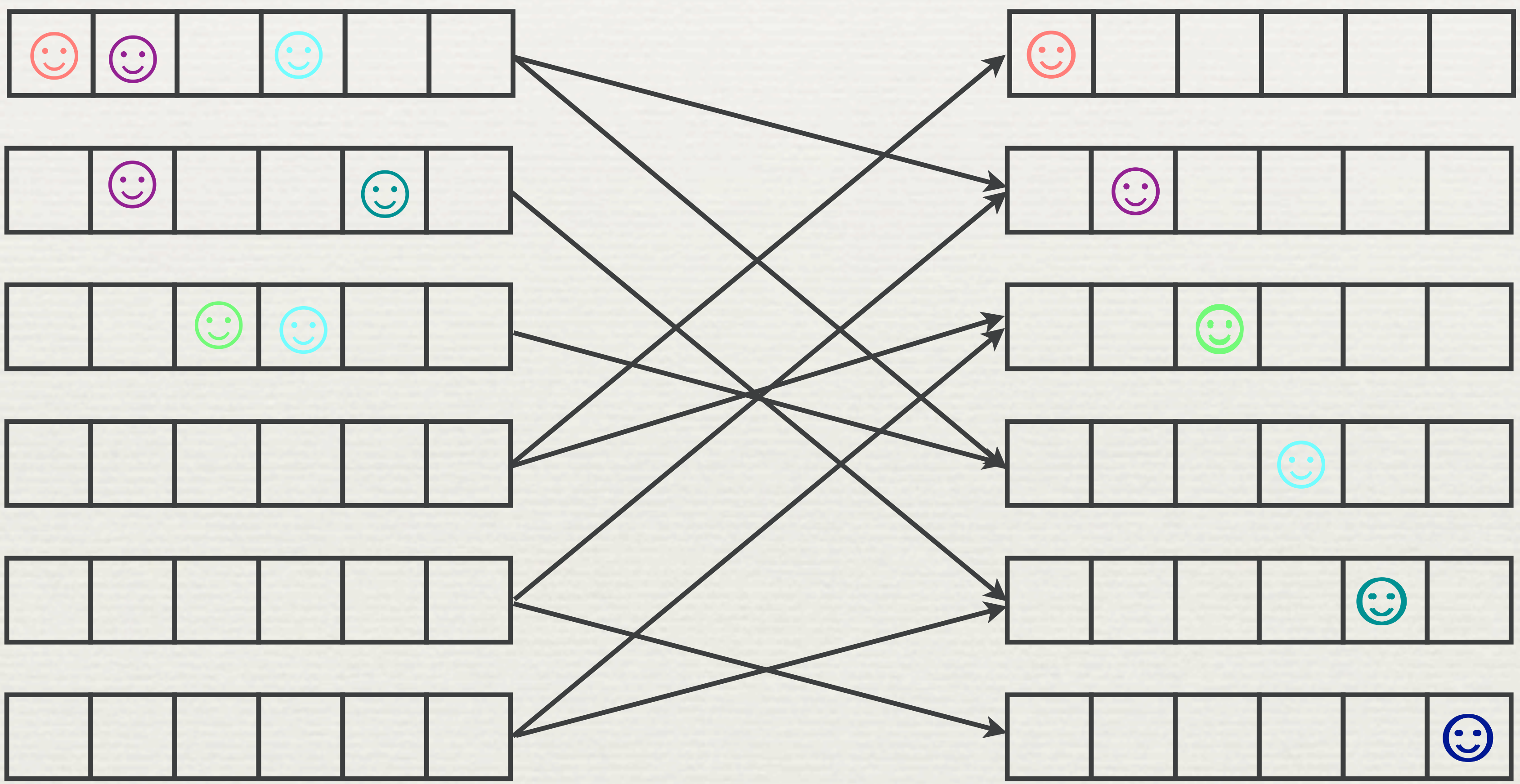
A round of updates



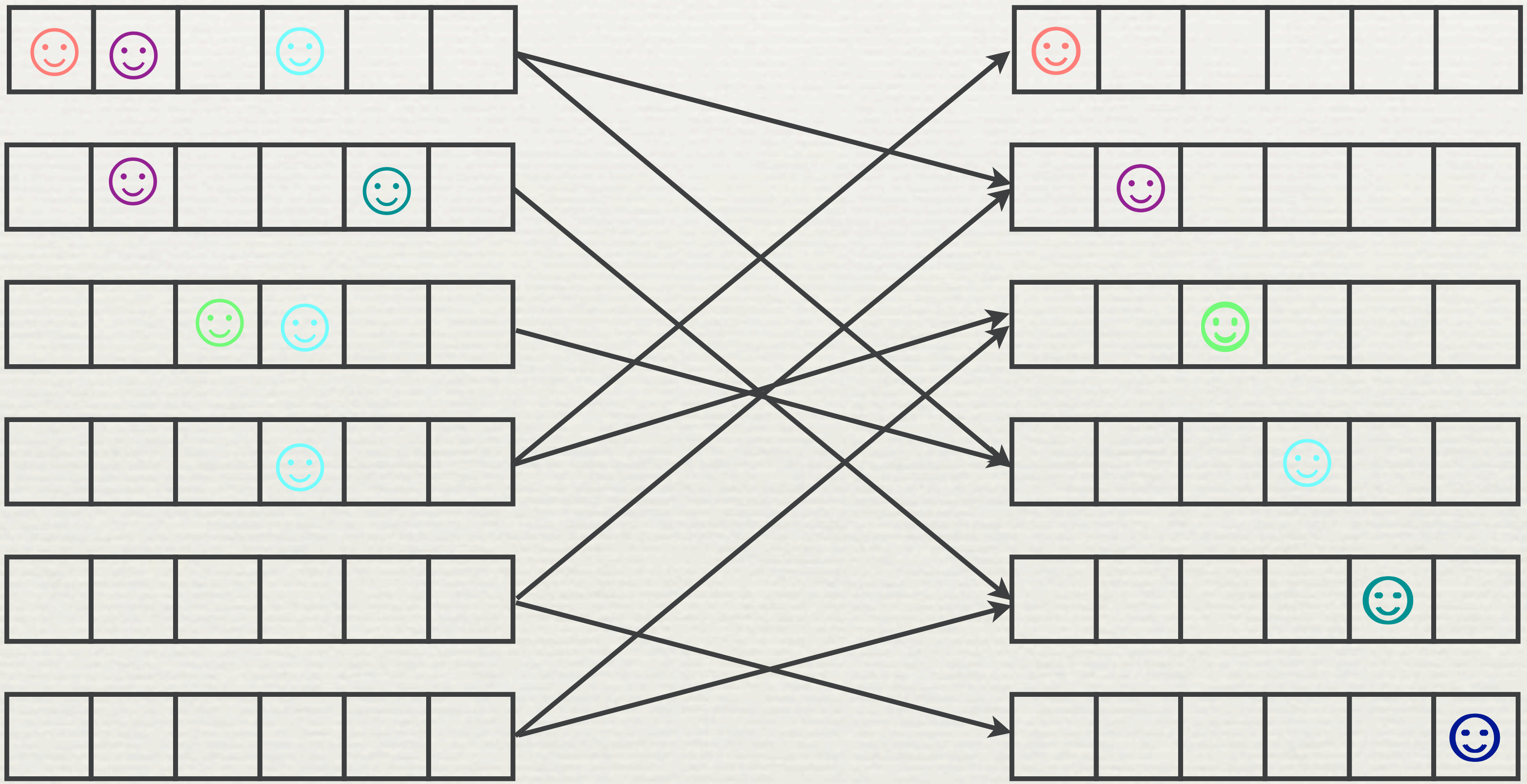
A round of updates



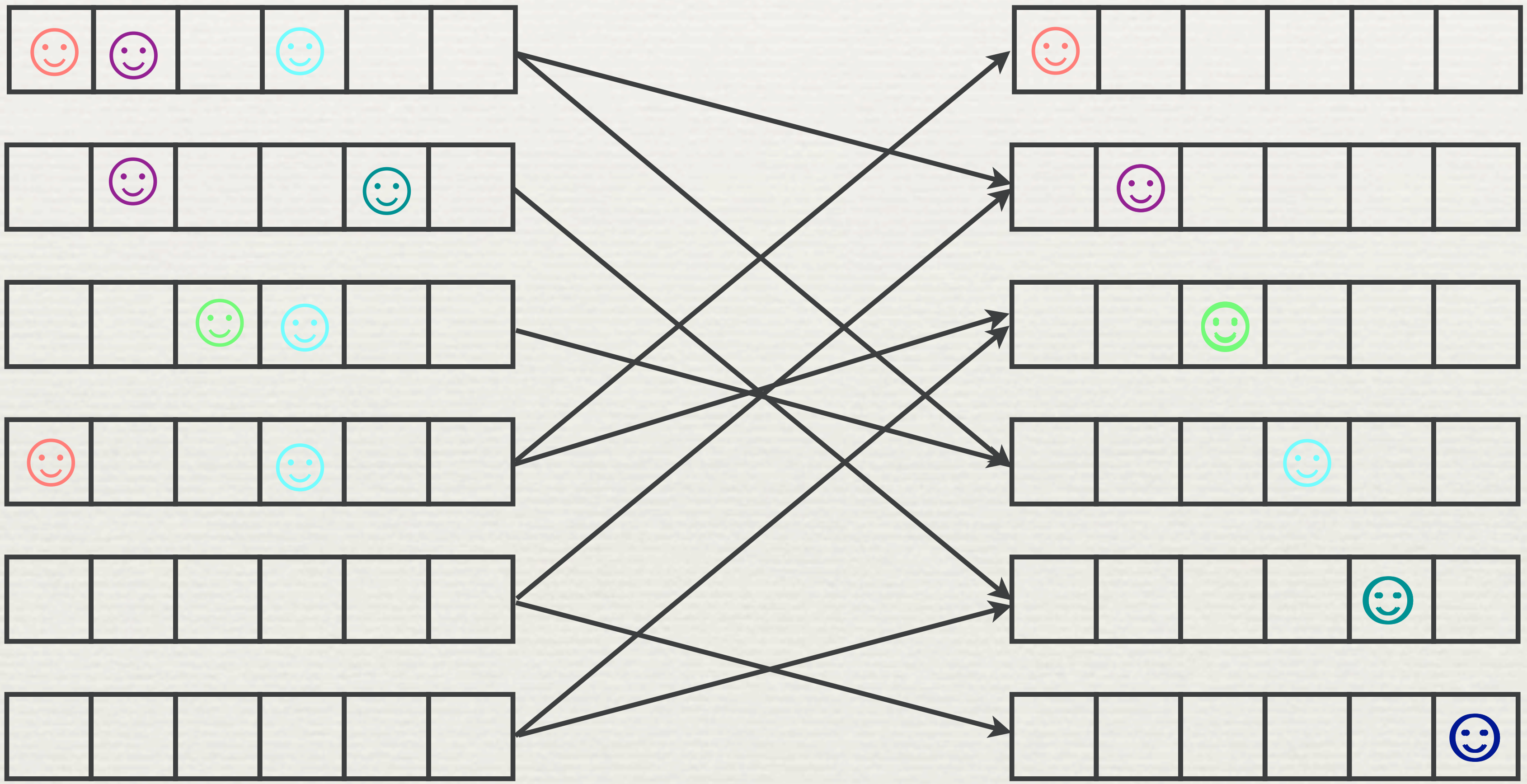
A round of updates



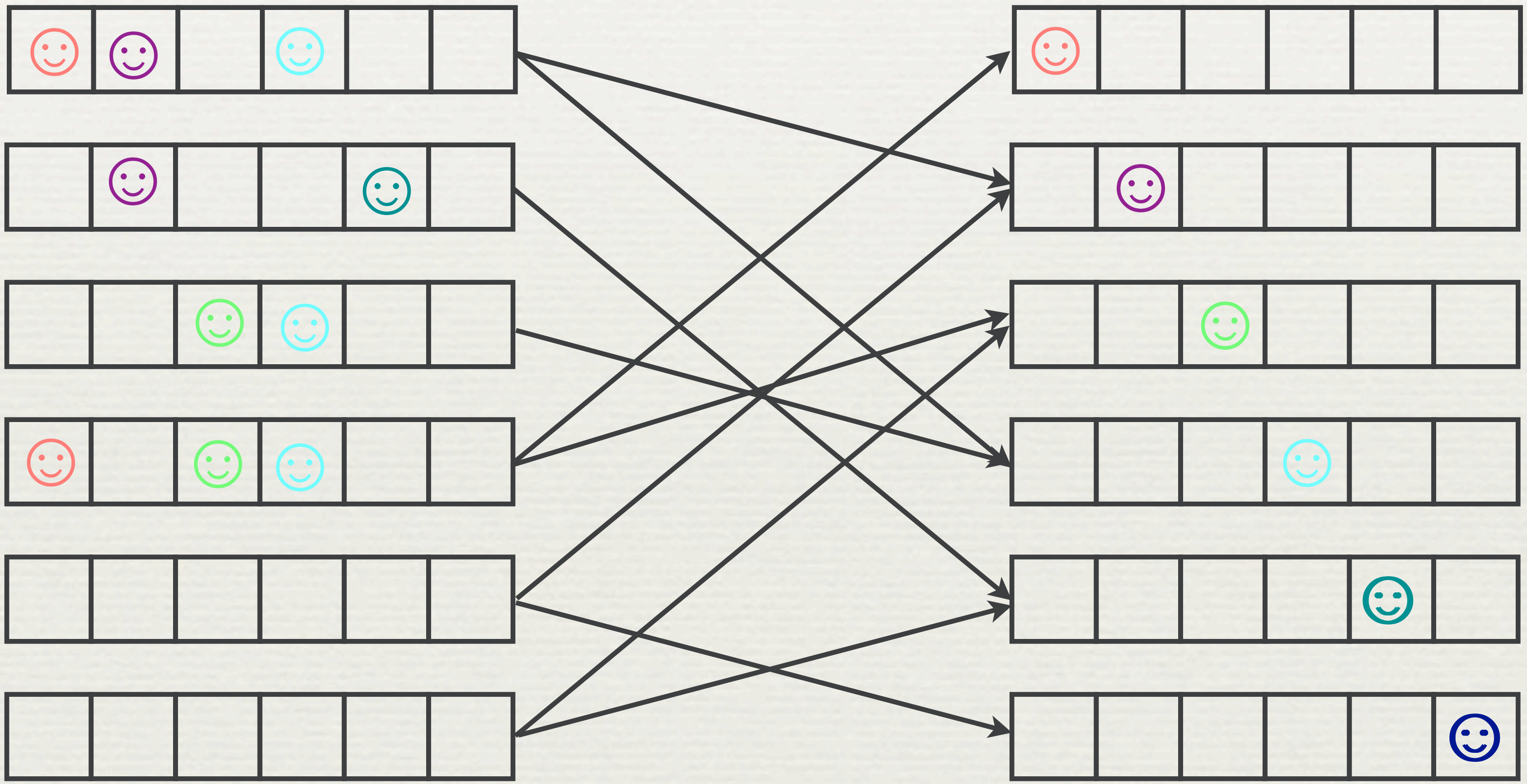
A round of updates



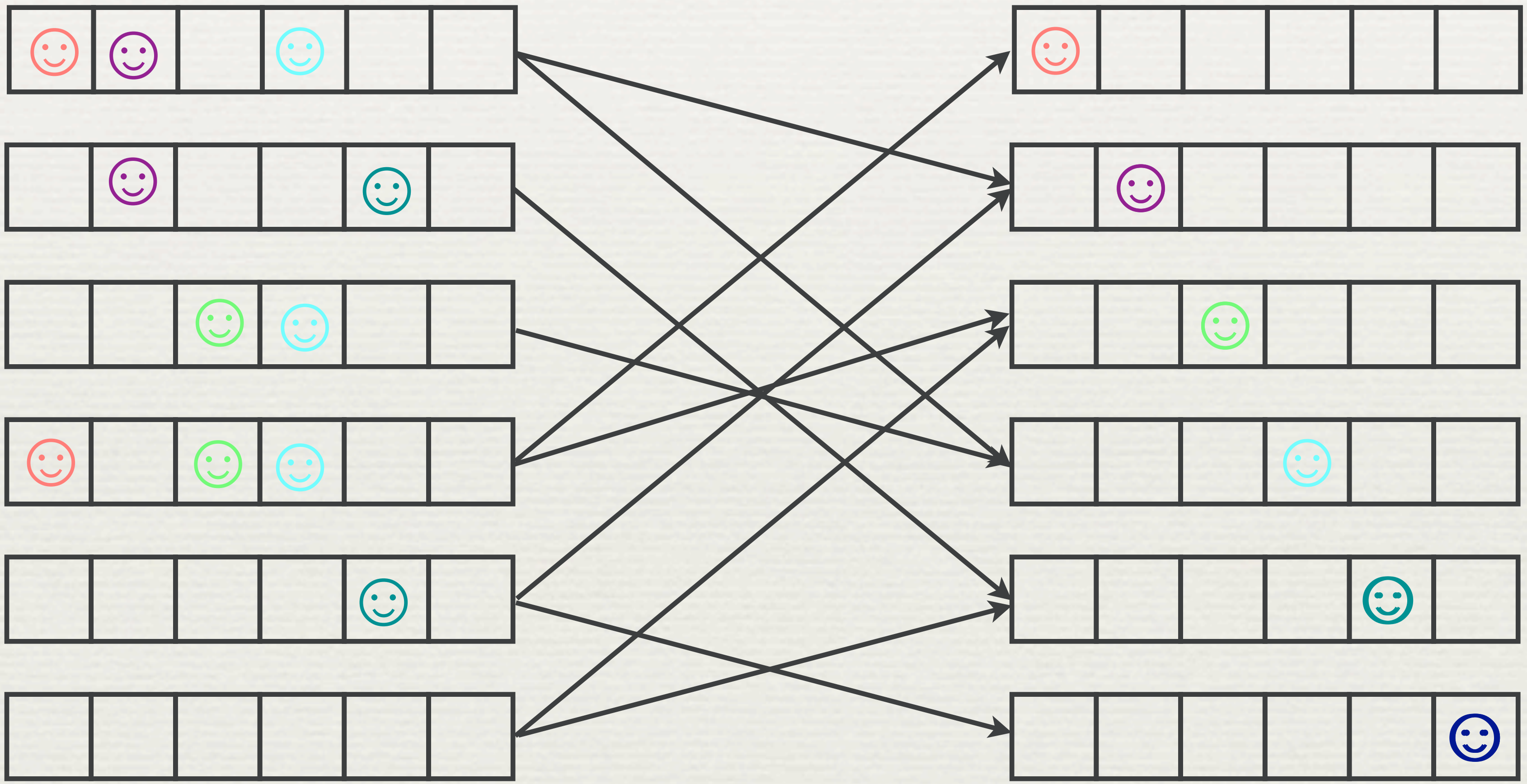
A round of updates



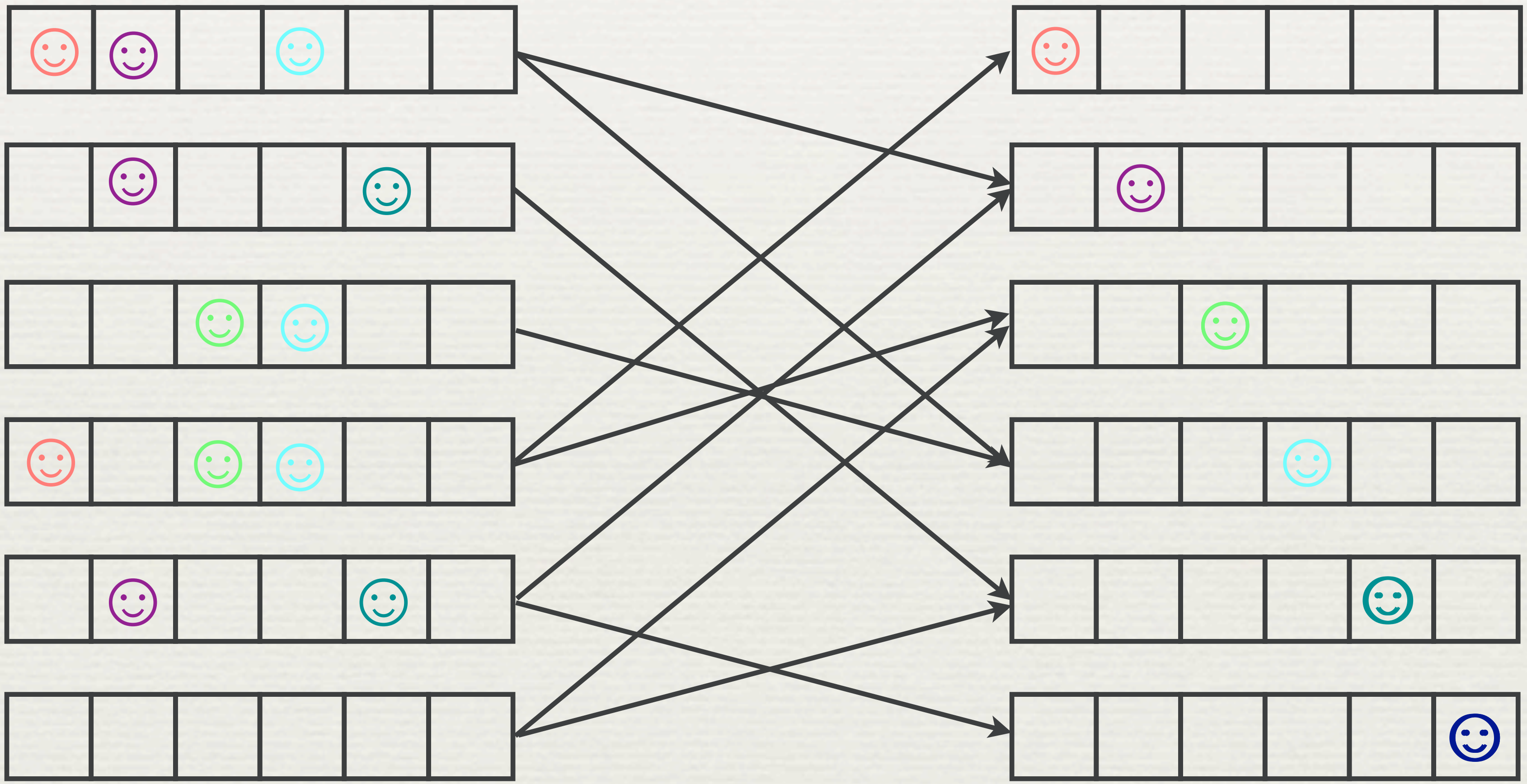
A round of updates



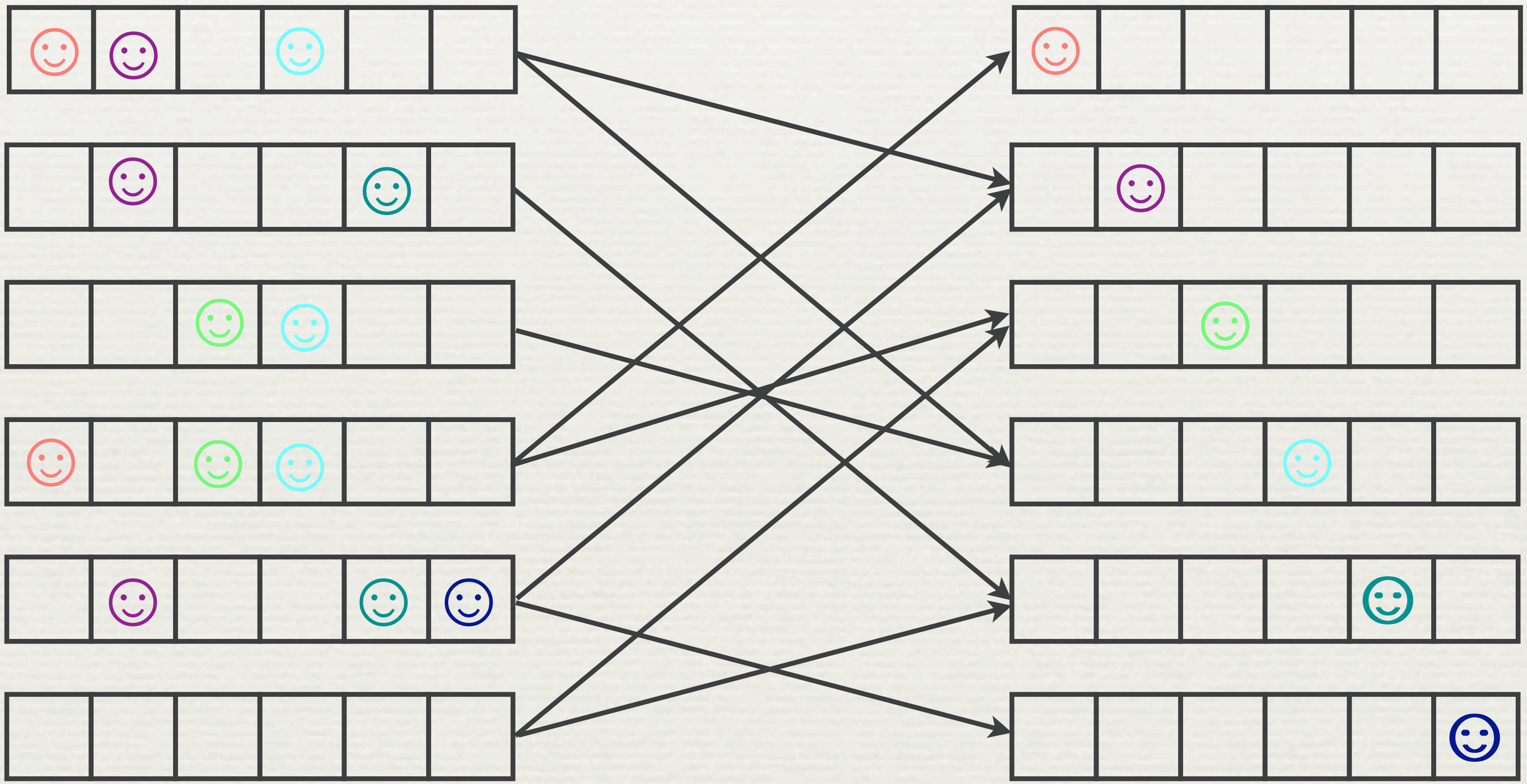
A round of updates



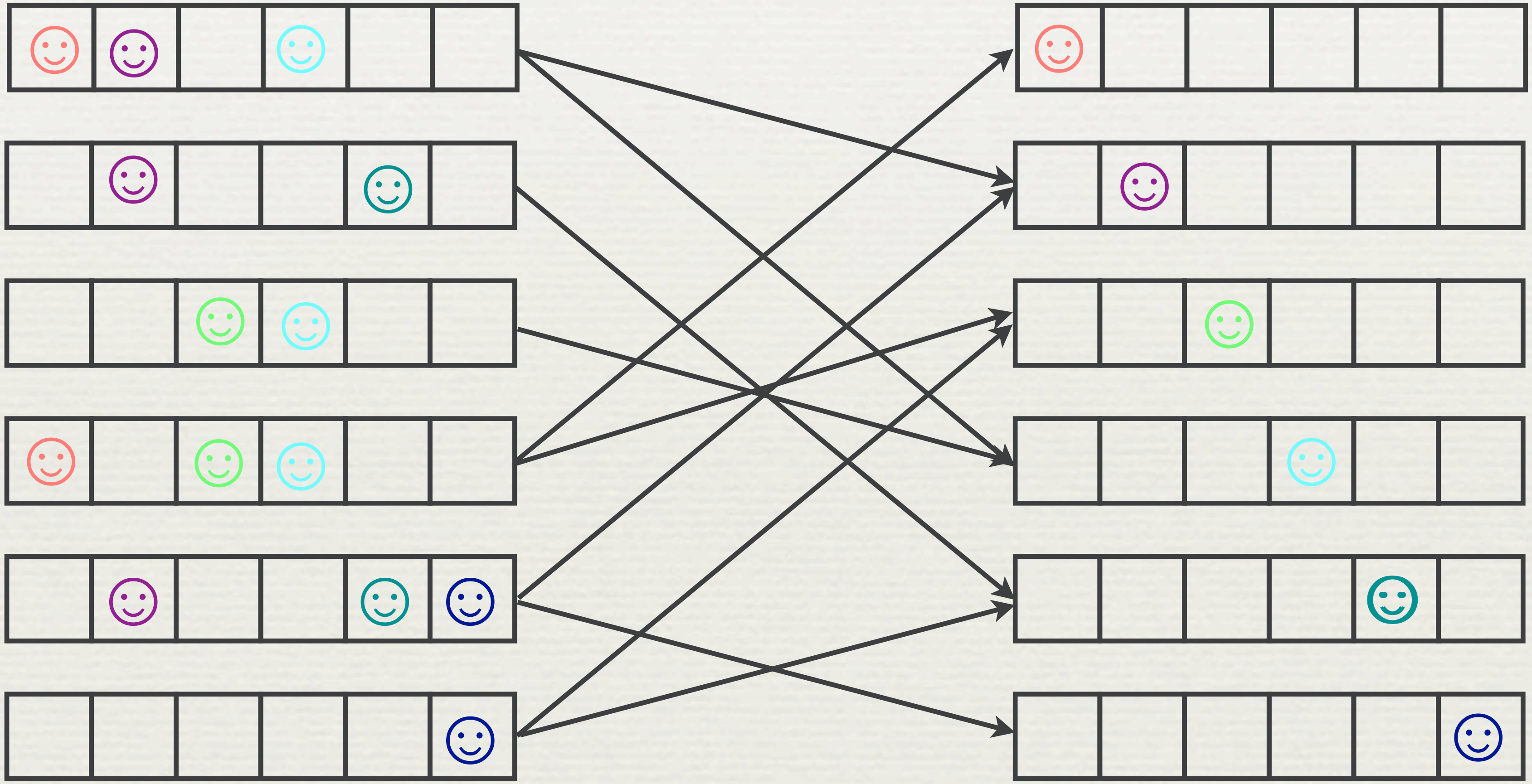
A round of updates



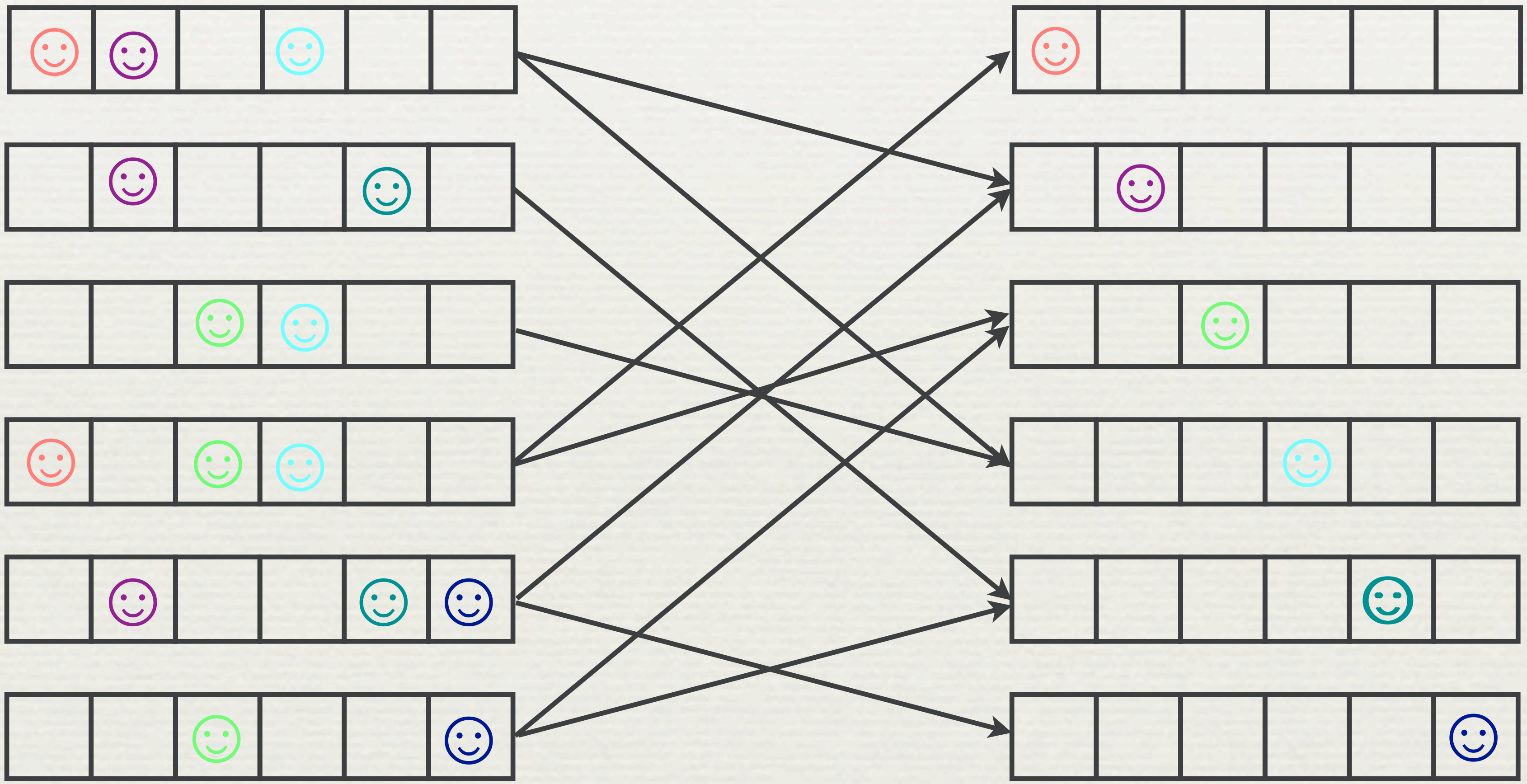
A round of updates



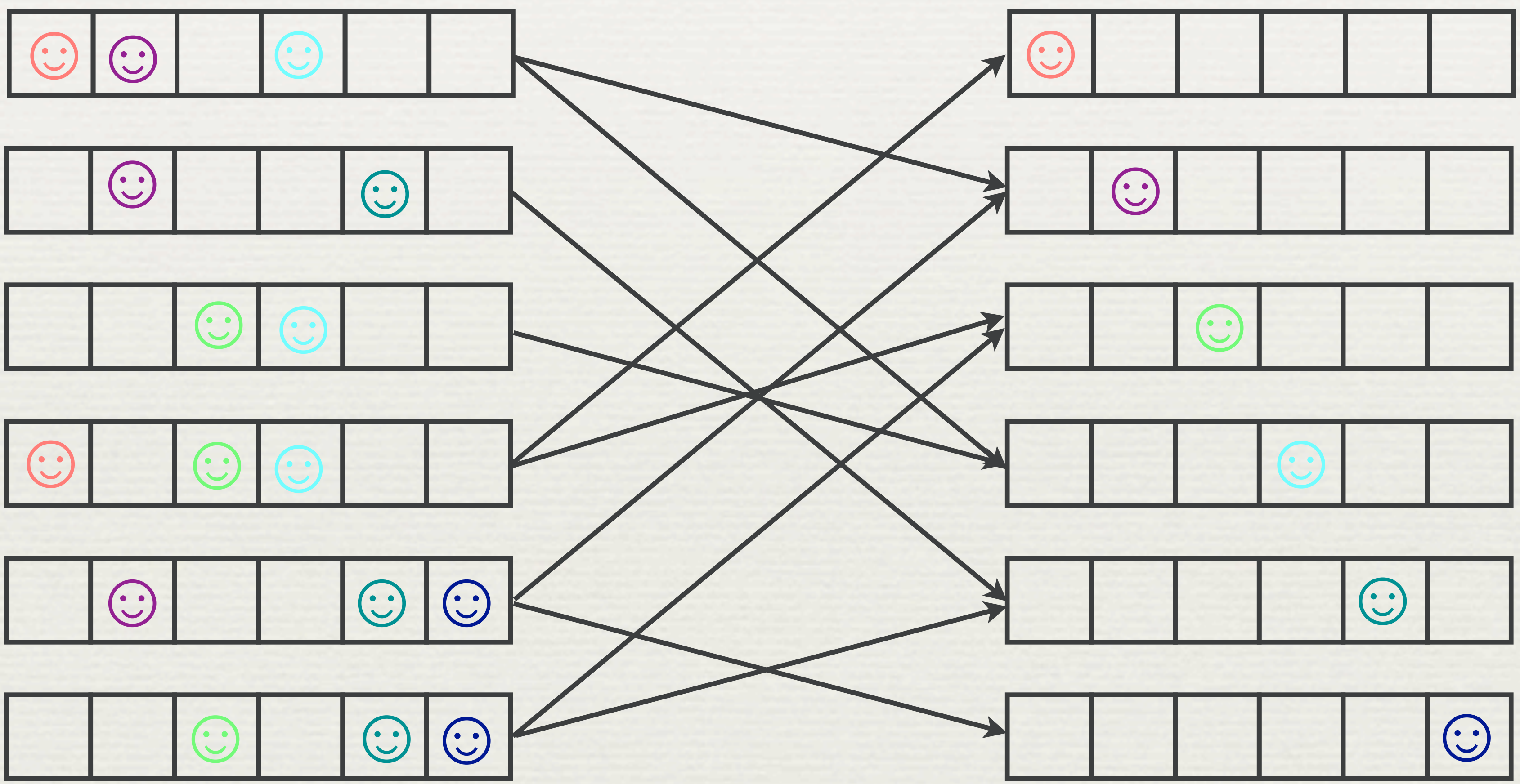
A round of updates



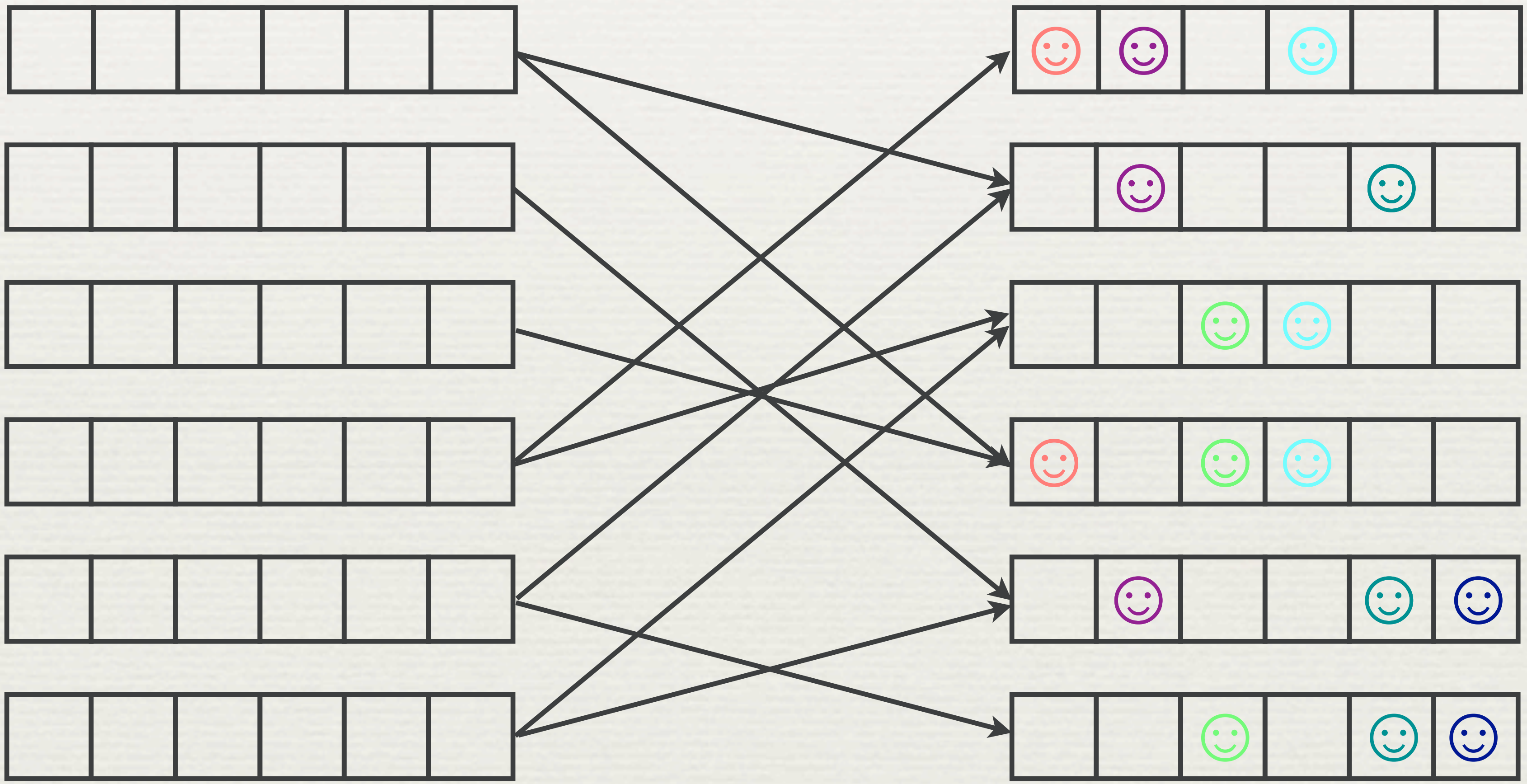
A round of updates



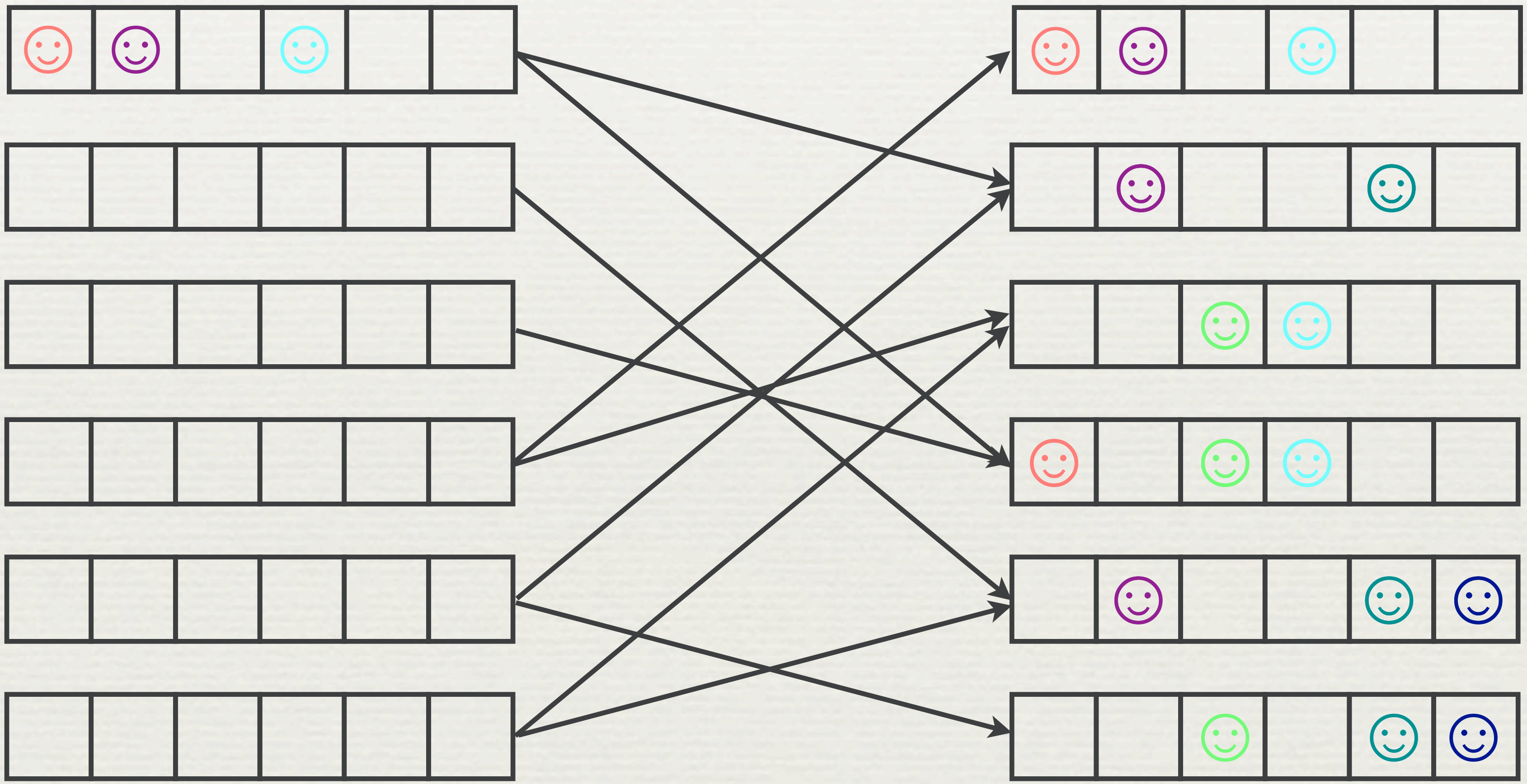
A round of updates



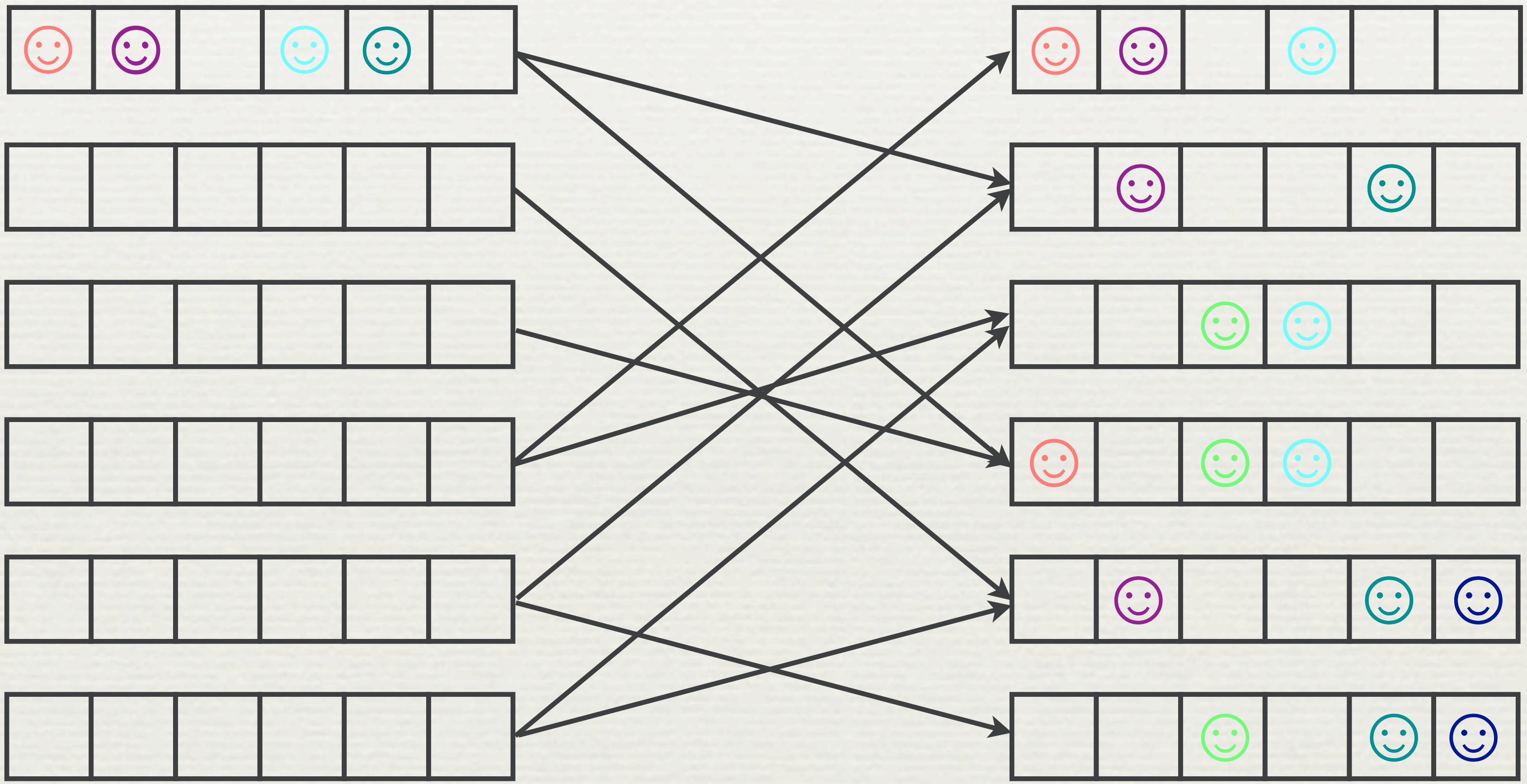
Another round...



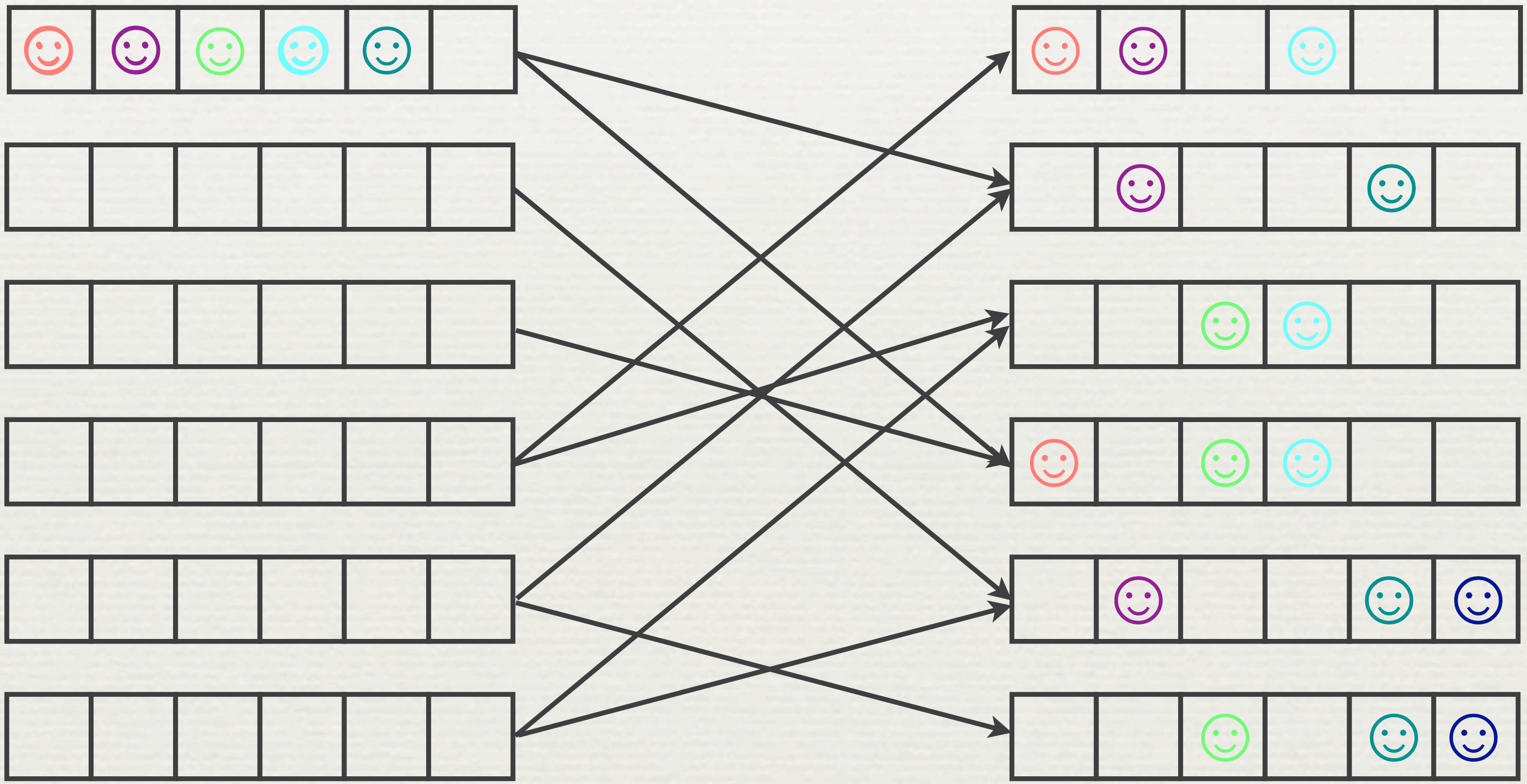
Another round...



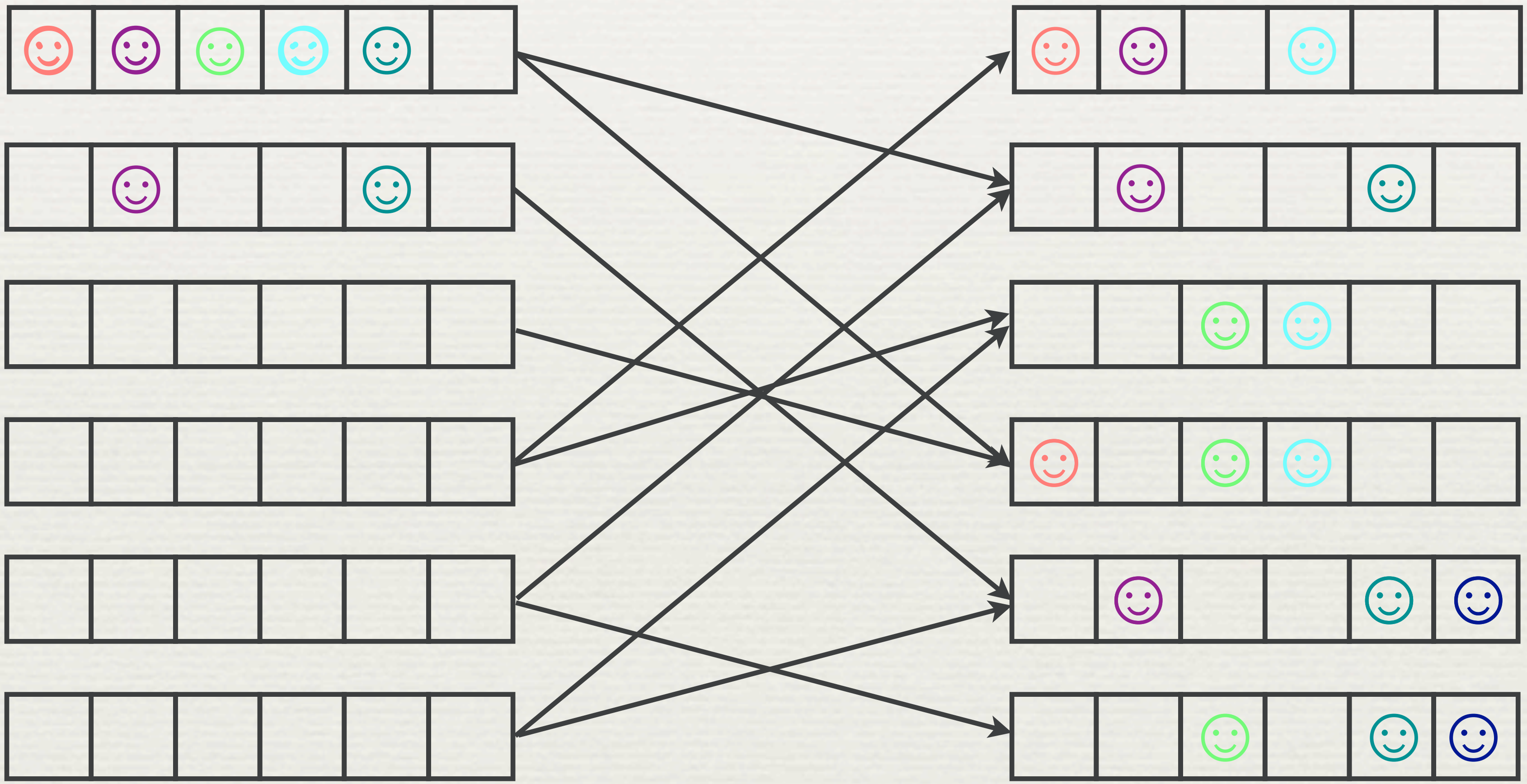
Another round...



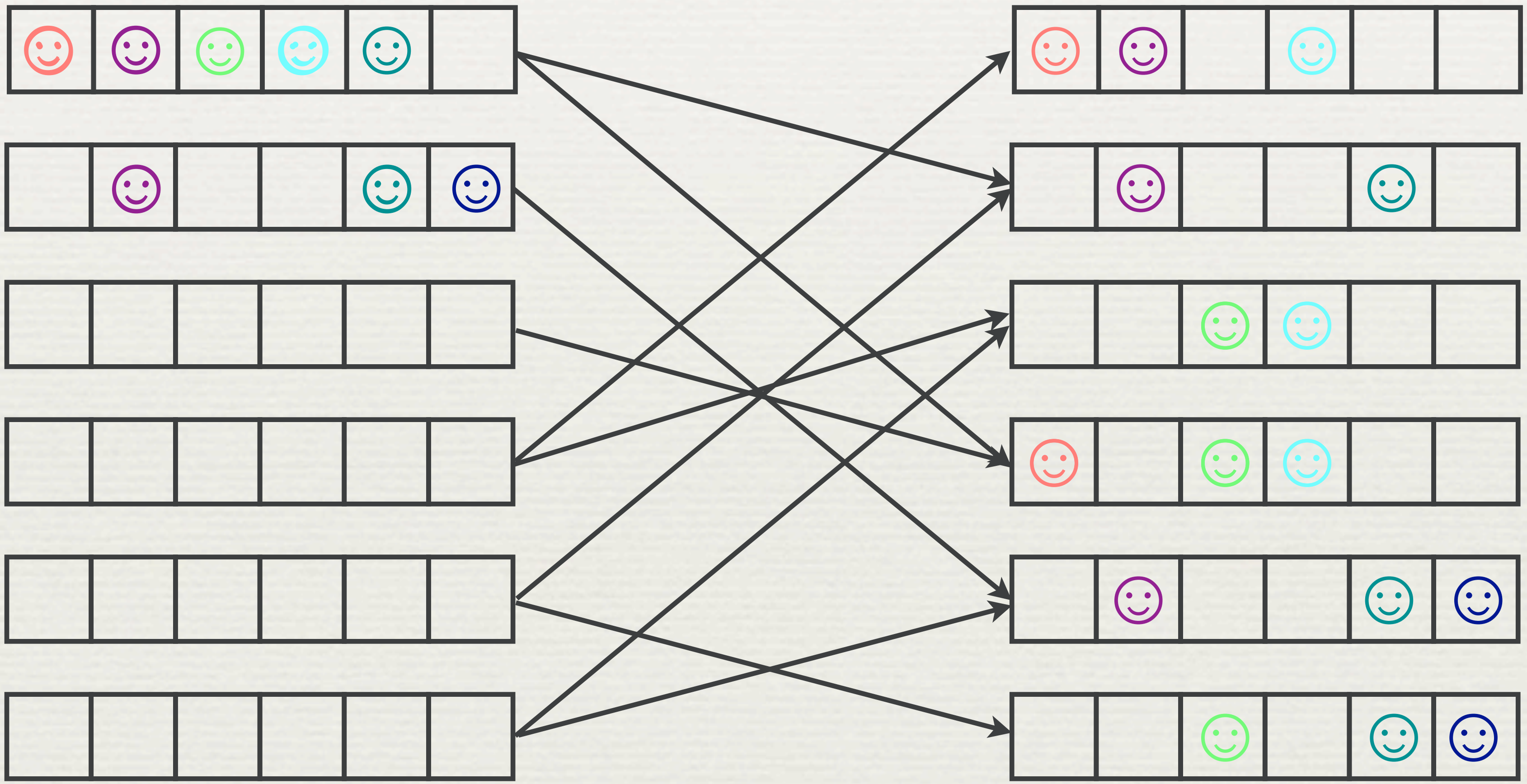
Another round...



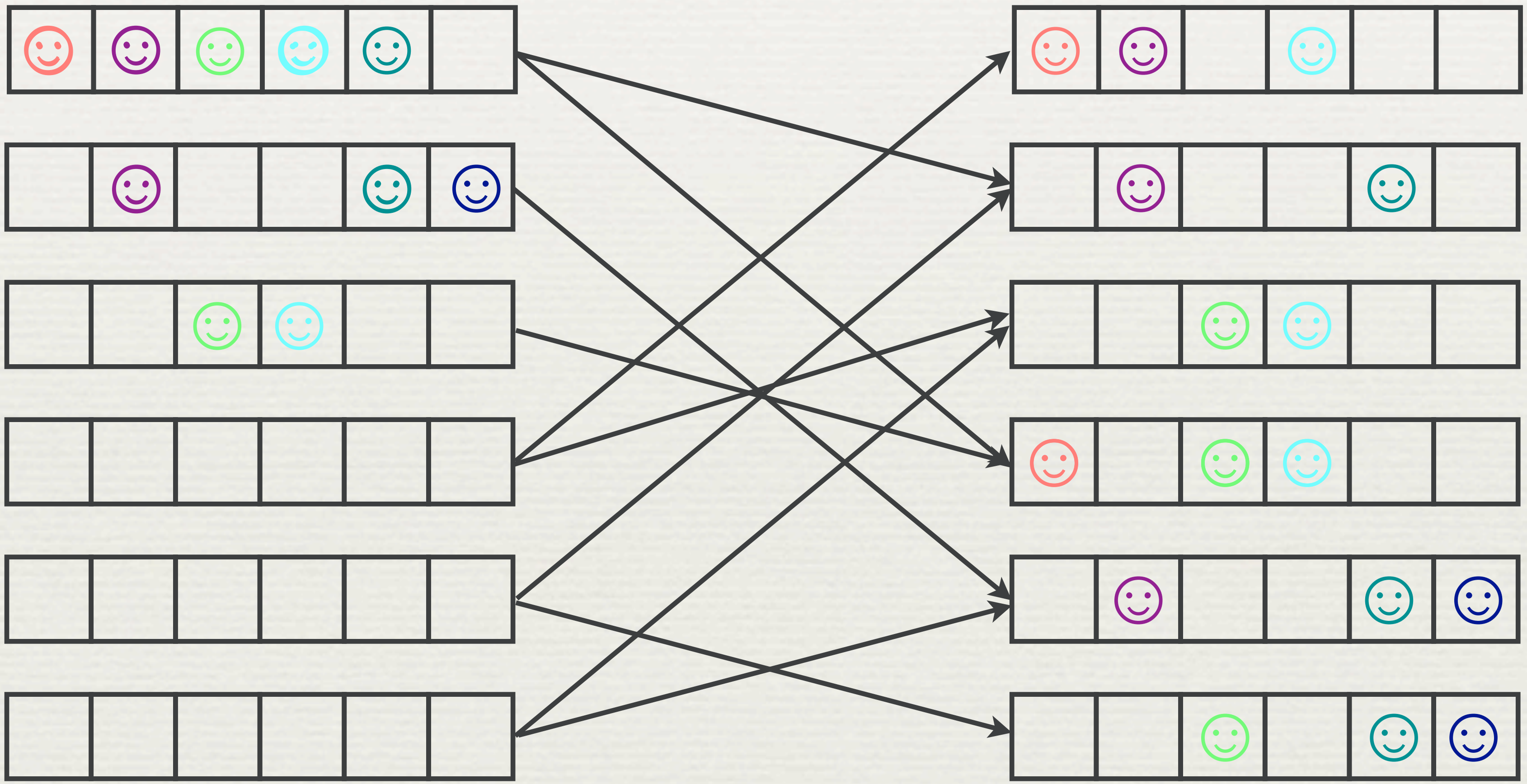
Another round...



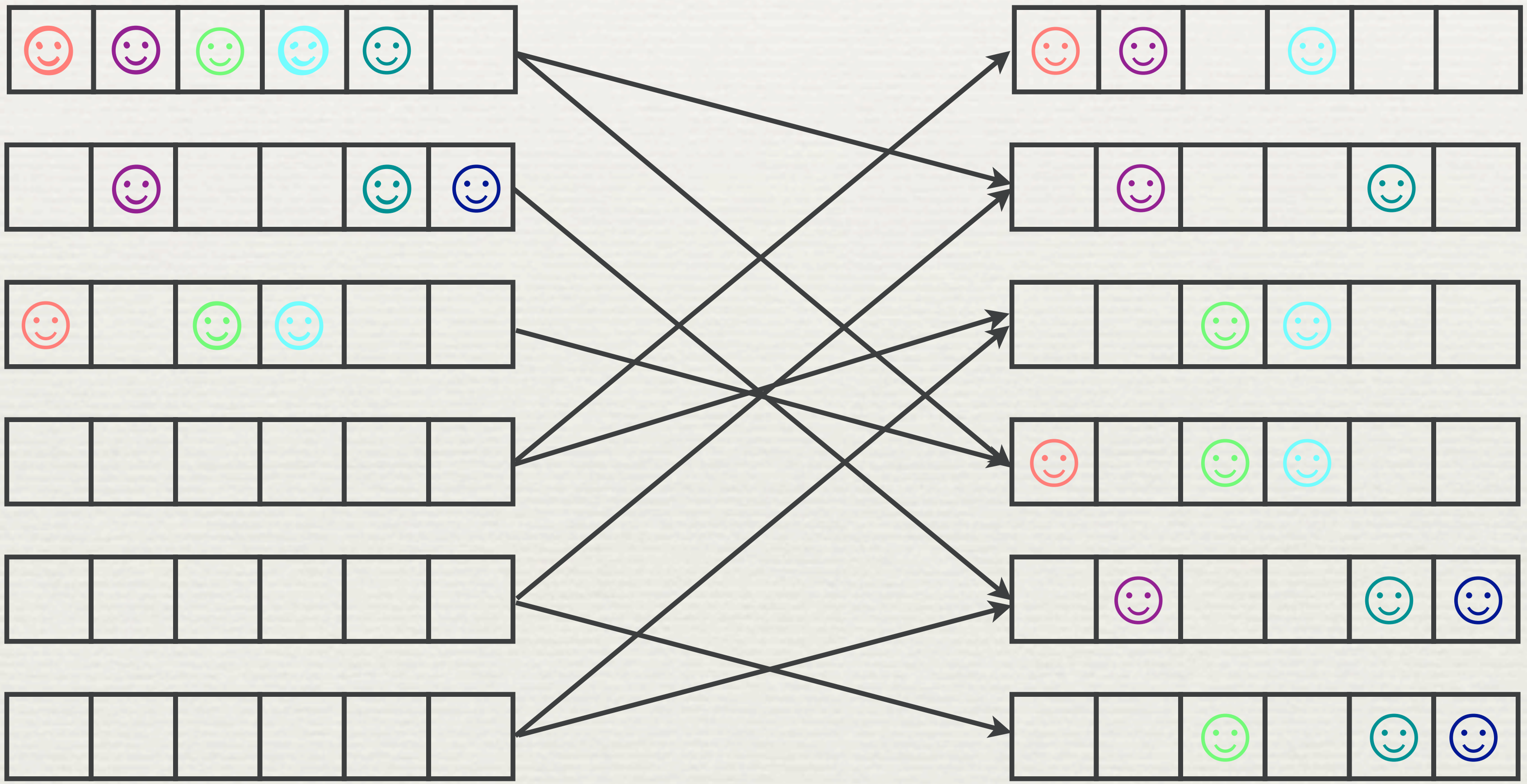
Another round...



Another round...



Another round...



Easy but expensive

Easy but expensive

- Each set uses linear space; overall quadratic

Easy but expensive

- Each set uses linear space; overall quadratic
- Impossible!

Easy but expensive

- Each set uses linear space; overall quadratic
- Impossible!
- But what if we use *approximate* sets?

Easy but expensive

- Each set uses linear space; overall quadratic
- Impossible!
- But what if we use *approximate* sets?
- Idea: use *probabilistic counters*, which represent sets but answer just to “size?” questions

Easy but expensive

- Each set uses linear space; overall quadratic
- Impossible!
- But what if we use *approximate* sets?
- Idea: use *probabilistic counters*, which represent sets but answer just to “size?” questions
- Very small!

Main trick

Main trick

- Choose an approximate set such that unions can be computed quickly

Main trick

- Choose an approximate set such that unions can be computed quickly
- ANF [Palmer *et al.*, KDD '02] uses Martin–Flajolet (MF) counters (log n+c space)

Main trick

- Choose an approximate set such that unions can be computed quickly
- ANF [Palmer *et al.*, KDD '02] uses Martin–Flajolet (MF) counters (log n+c space)
- We use HyperLogLog counters [Flajolet *et al.*, 2007] (loglog n space)

Main trick

- Choose an approximate set such that unions can be computed quickly
- ANF [Palmer *et al.*, KDD '02] uses Martin–Flajolet (MF) counters (log n+c space)
- We use HyperLogLog counters [Flajolet *et al.*, 2007] (loglog n space)
- MF counters can be combined with an OR

Main trick

- Choose an approximate set such that unions can be computed quickly
- ANF [Palmer *et al.*, KDD '02] uses Martin–Flajolet (MF) counters (log *n*+*c* space)
- We use HyperLogLog counters [Flajolet *et al.*, 2007] (loglog *n* space)
- MF counters can be combined with an OR
- We use *broadword programming* to combine HyperLogLog counters quickly!

HyperLogLog counters

HyperLogLog counters

- Instead of actually counting, we *observe* a statistical feature of a set (think stream) of elements

HyperLogLog counters

- Instead of actually counting, we *observe* a statistical feature of a set (think stream) of elements
- The feature: the number of trailing zeroes of the value of a **very good** hash function

HyperLogLog counters

- Instead of actually counting, we *observe* a statistical feature of a set (think stream) of elements
- The feature: the number of trailing zeroes of the value of a **very good** hash function
- We keep track of the maximum m (log log n bits!)

HyperLogLog counters

- Instead of actually counting, we *observe* a statistical feature of a set (think stream) of elements
- The feature: the number of trailing zeroes of the value of a **very good** hash function
- We keep track of the maximum m (log log n bits!)
- The number of distinct elements $\propto 2^m$

HyperLogLog counters

- Instead of actually counting, we *observe* a statistical feature of a set (think stream) of elements
- The feature: the number of trailing zeroes of the value of a **very good** hash function
- We keep track of the maximum m (log log n bits!)
- The number of distinct elements $\propto 2^m$
- **Important:** the counter of stream AB is simply the maximum of the counters of A and B!

Many, many counters...

Many, many counters...

- To increase confidence, we need *several* counters (usually 2^b, $b \geq 4$) and take their harmonic mean

Many, many counters...

- To increase confidence, we need *several* counters (usually 2^b, $b \geq 4$) and take their harmonic mean
- Thus each set is represented by a list of small (typically 5-bit) counters (unlikely >7 bits!)

Many, many counters...

- To increase confidence, we need *several* counters (usually 2^b, $b \geq 4$) and take their harmonic mean
- Thus each set is represented by a list of small (typically 5-bit) counters (unlikely >7 bits!)
- To compute the union of two sets these must be maximized one-by-one

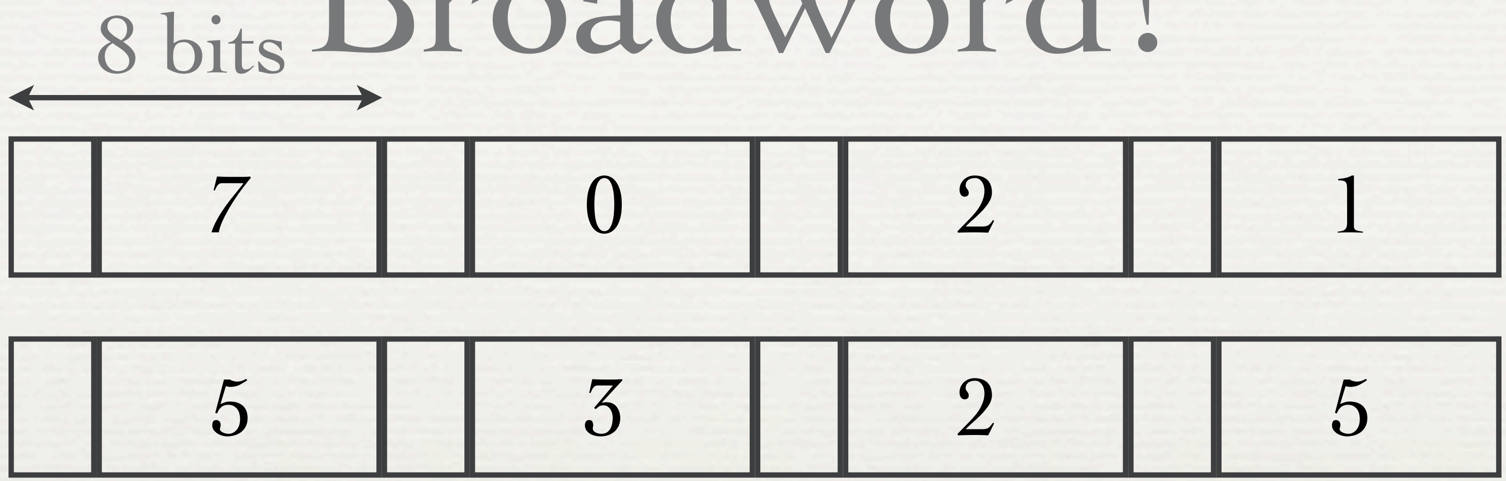
Many, many counters...

- To increase confidence, we need *several* counters (usually 2^b, $b \geq 4$) and take their harmonic mean
- Thus each set is represented by a list of small (typically 5-bit) counters (unlikely >7 bits!)
- To compute the union of two sets these must be maximized one-by-one
- Extracting by shifts, maximizing and putting back by shifts is unbearably slow

Many, many counters...

- To increase confidence, we need *several* counters (usually 2^b, $b \geq 4$) and take their harmonic mean
- Thus each set is represented by a list of small (typically 5-bit) counters (unlikely >7 bits!)
- To compute the union of two sets these must be maximized one-by-one
- Extracting by shifts, maximizing and putting back by shifts is unbearably slow
- In the Martin–Flajolet case just OR the features!

Broadword!



Broadword!

8 bits

1	7	1	0	1	2	1	1

0	5	0	3	0	2	0	5

Broadword!

8 bits

1	7	1	0	1	2	1	1

$-$

0	5	0	3	0	2	0	5

$=$

Broadword!

8 bits

1	7	1	0	1	2	1	1

−

0	5	0	3	0	2	0	5

=

	2		125		0		124

Broadword!

8 bits

1	7	1	0	1	2	1	1

−

0	5	0	3	0	2	0	5

=

1	2	0	125	1	0	0	124

Broadword!

8 bits

1	7	1	0	1	2	1	1

-

0	5	0	3	0	2	0	5

=

1	2	0	125	1	0	0	124

1		1		1		1	

Broadword!

8 bits

1	7	1	0	1	2	1	1

-

0	5	0	3	0	2	0	5

=

	2		125		0		124

1	1	1	0	1	1	1	0

Broadword!

8 bits

1	7	1	0	1	2	1	1

~

0	5	0	3	0	2	0	5

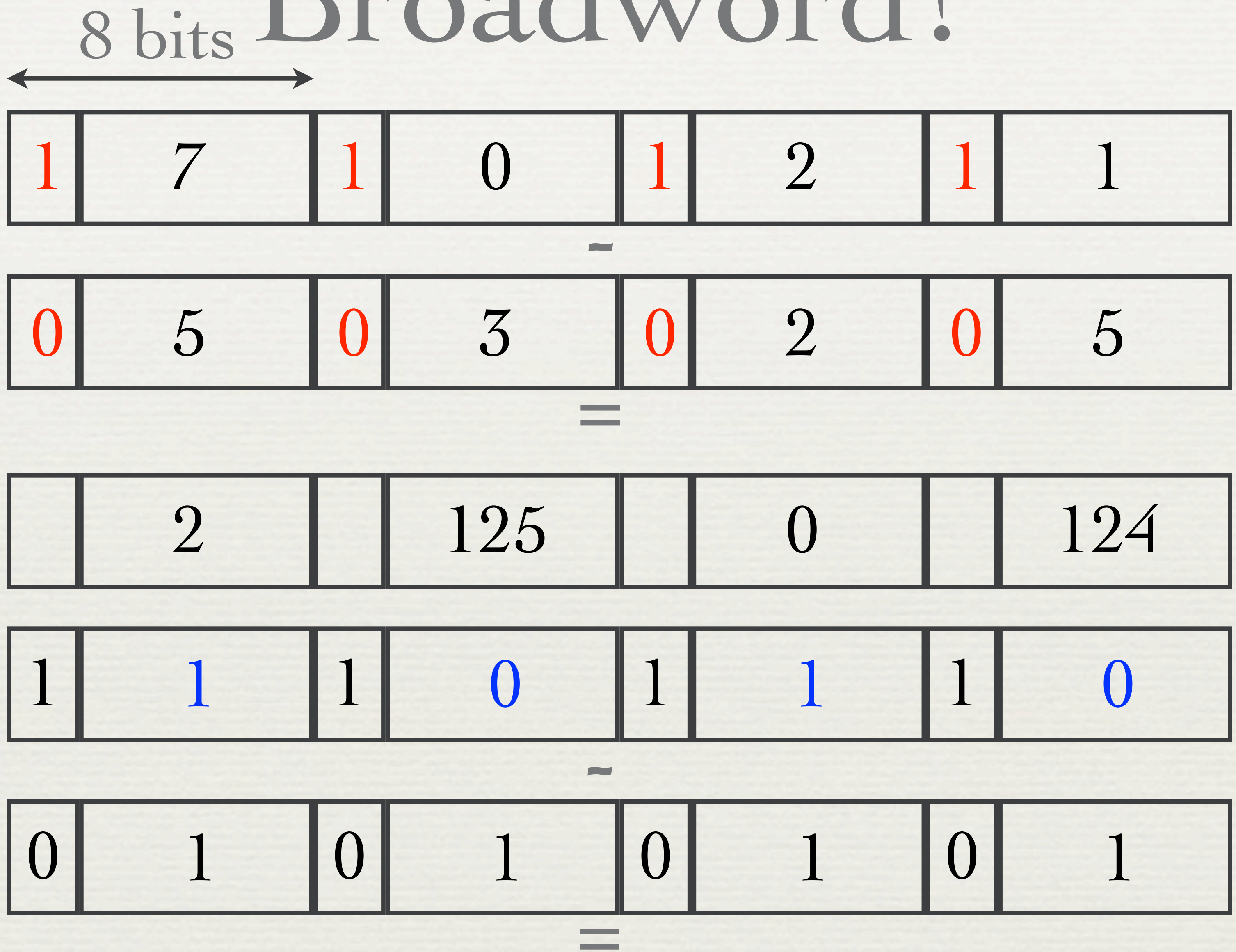
=

	2		125		0		124

1	1	1	0	1	1	1	0

0	1	0	1	0	1	0	1

Broadword!



Broadword!

8 bits

1	7	1	0	1	2	1	1

−

0	5	0	3	0	2	0	5

=

	2		125		0		124

1	1	1	0	1	1	1	0

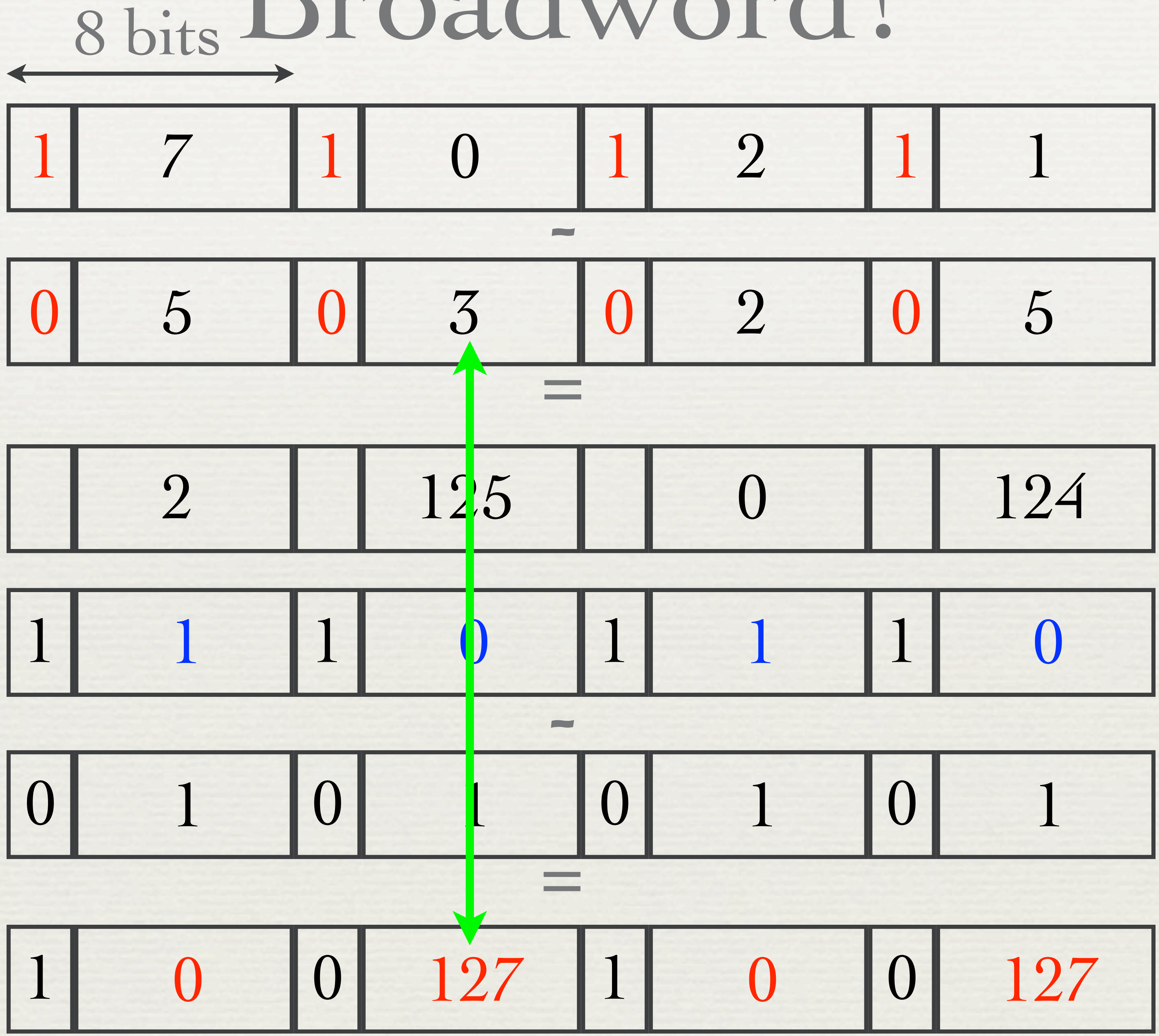
−

0	1	0	1	0	1	0	1

=

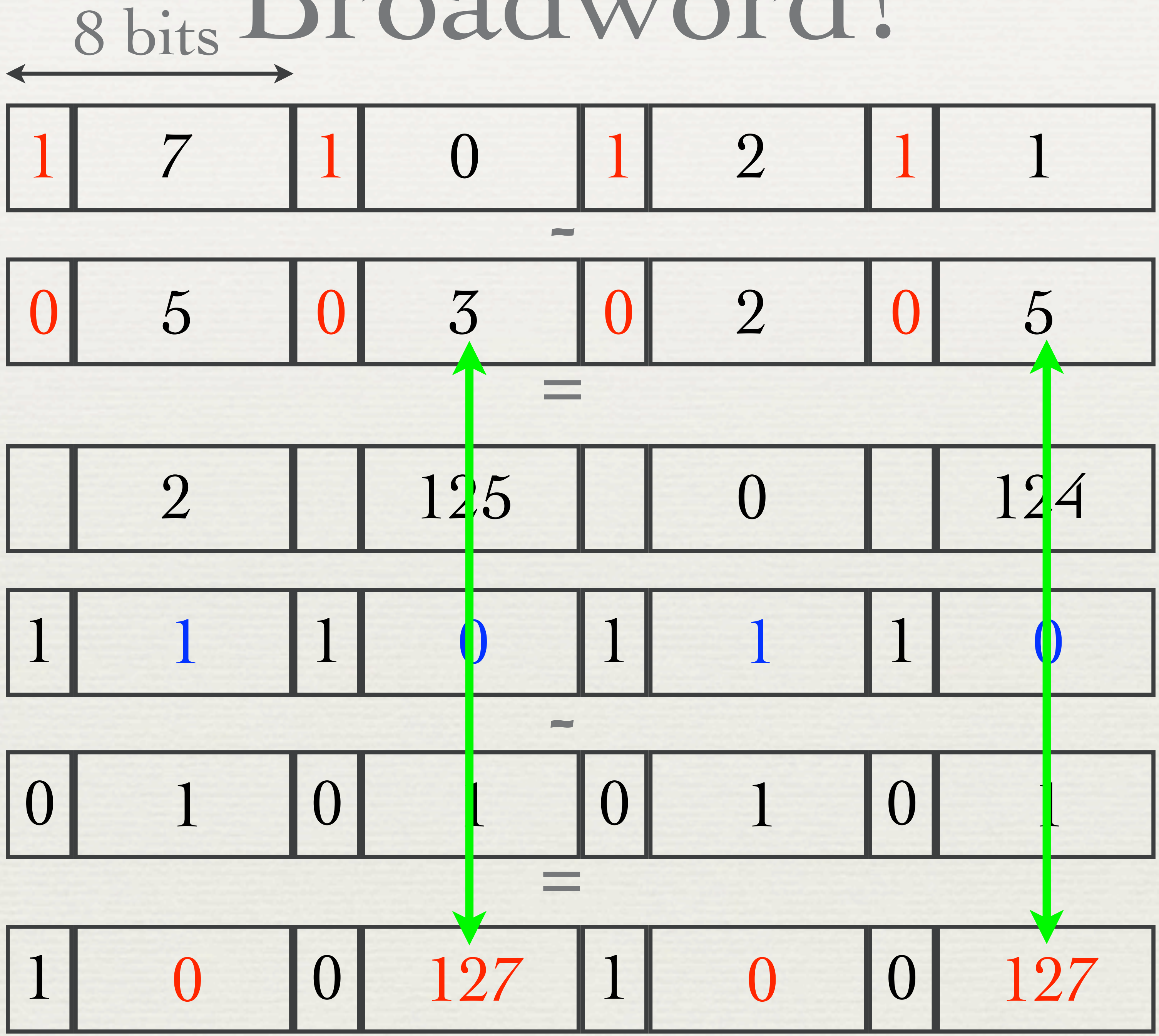
1	0	0	127	1	0	0	127

Broadword!

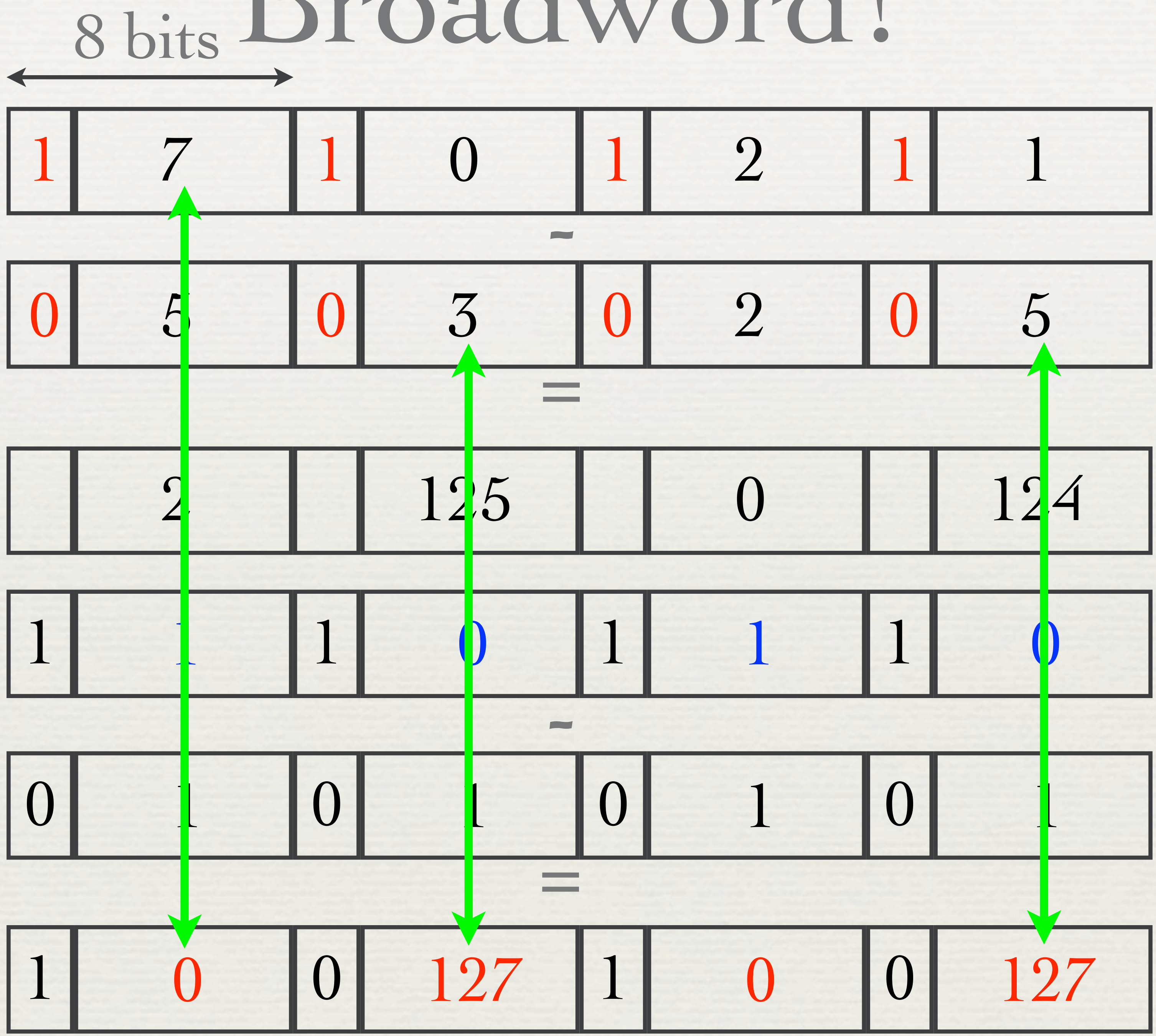


8 bits
1 7 1 0 1 2 1 1
-
0 5 0 3 0 2 0 5
=
2 125 0 124
1 1 1 0 1 1 1 0
-
0 1 0 1 0 1 0 1
=
1 0 0 127 1 0 0 127

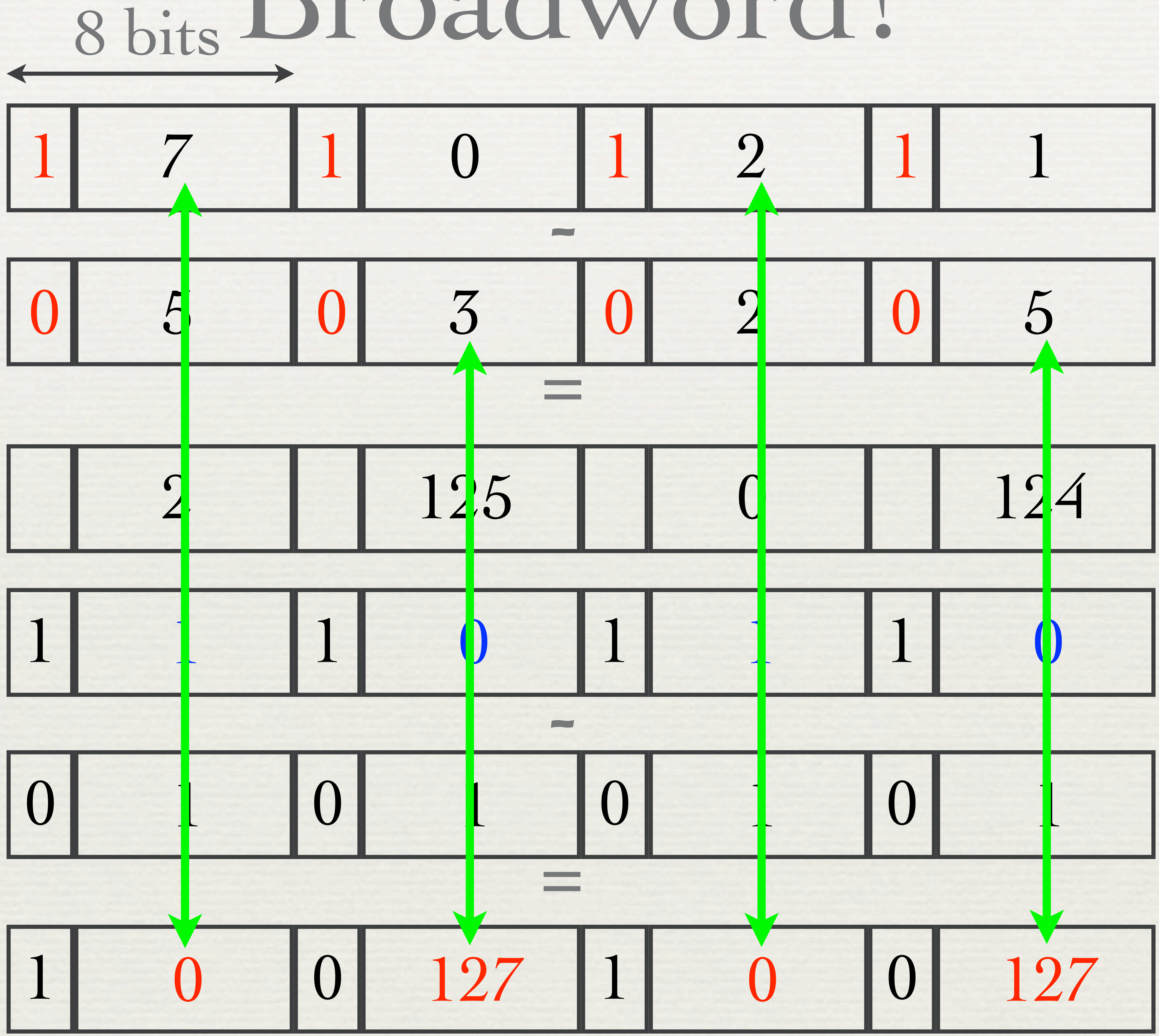
Broadword!


8 bits
1 7 1 0 1 2 1 1
-
0 5 0 3 0 2 0 5
=
2 125 0 124
1 1 1 0 1 1 1 0
-
0 1 0 1 0 1 0 1
=
1 0 0 127 1 0 0 127

Broadword!


8 bits
1 7 1 0 1 2 1 1
-
0 5 0 3 0 2 0 5
=
2 125 0 124
1 1 1 0 1 1 1 0
-
0 1 0 1 0 1 0 1
=
1 0 0 127 1 0 0 127

Broadword!



8 bits
1 7 1 0 1 2 1 1
-
0 5 0 3 0 2 0 5
=
2 125 0 124
1 1 1 0 1 1 1 0
-
0 1 0 1 0 1 0 1
=
1 0 0 127 1 0 0 127

Real speed?

Real speed?

- Large size: HADI {Kang et al., 2010} is a Hadoop-conscious implementation of ANF. Takes 30 minutes on a 200K-node graph (on one of the 50 world largest supercomputers). HyperANF does the same in 2.25min on our workstation (15 min on this laptop).

On Facebook?

On Facebook?

- When I presented HyperANF at WWW 2011, I suggested it would have been nice to run it on Facebook

On Facebook?

- When I presented HyperANF at WWW 2011, I suggested it would have been nice to run it on Facebook
- Lars Backstrom was there and said “why not”?

On Facebook?

- When I presented HyperANF at WWW 2011, I suggested it would have been nice to run it on Facebook
- Lars Backstrom was there and said “why not”?
- We started interacting few months after

On Facebook?

- When I presented HyperANF at WWW 2011, I suggested it would have been nice to run it on Facebook
- Lars Backstrom was there and said “why not”?
- We started interacting few months after
- No data moving: Java jars were sent from the LAW and run at facebook

On Facebook?

- When I presented HyperANF at WWW 2011, I suggested it would have been nice to run it on Facebook
- Lars Backstrom was there and said “why not”?
- We started interacting few months after
- No data moving: Java jars were sent from the LAW and run at facebook
- Quite crazy software management setup, believe me...

Experiments (time)

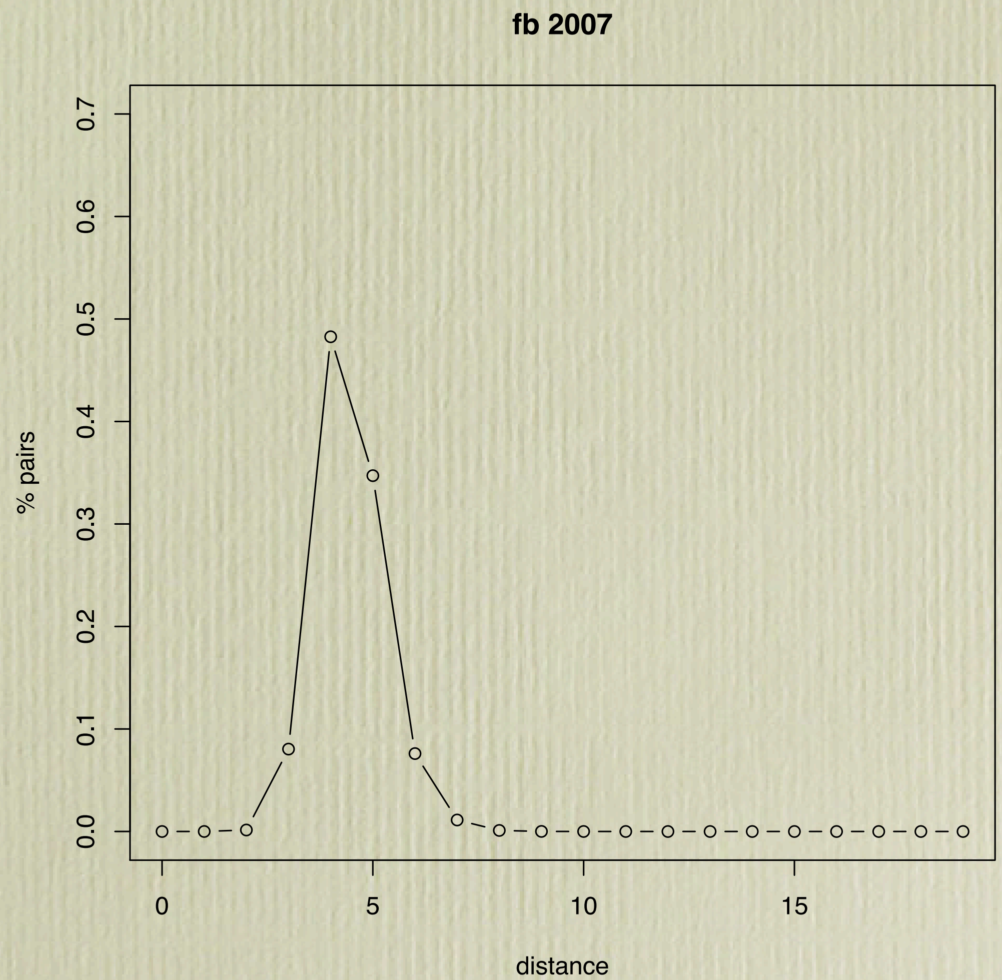
- We ran our experiments on snapshots of facebook
    - Jan 1, 2007
    - Jan 1, 2008 ...
    - Jan 1, 2011
    - {current} May, 2011

Experiments (dataset)

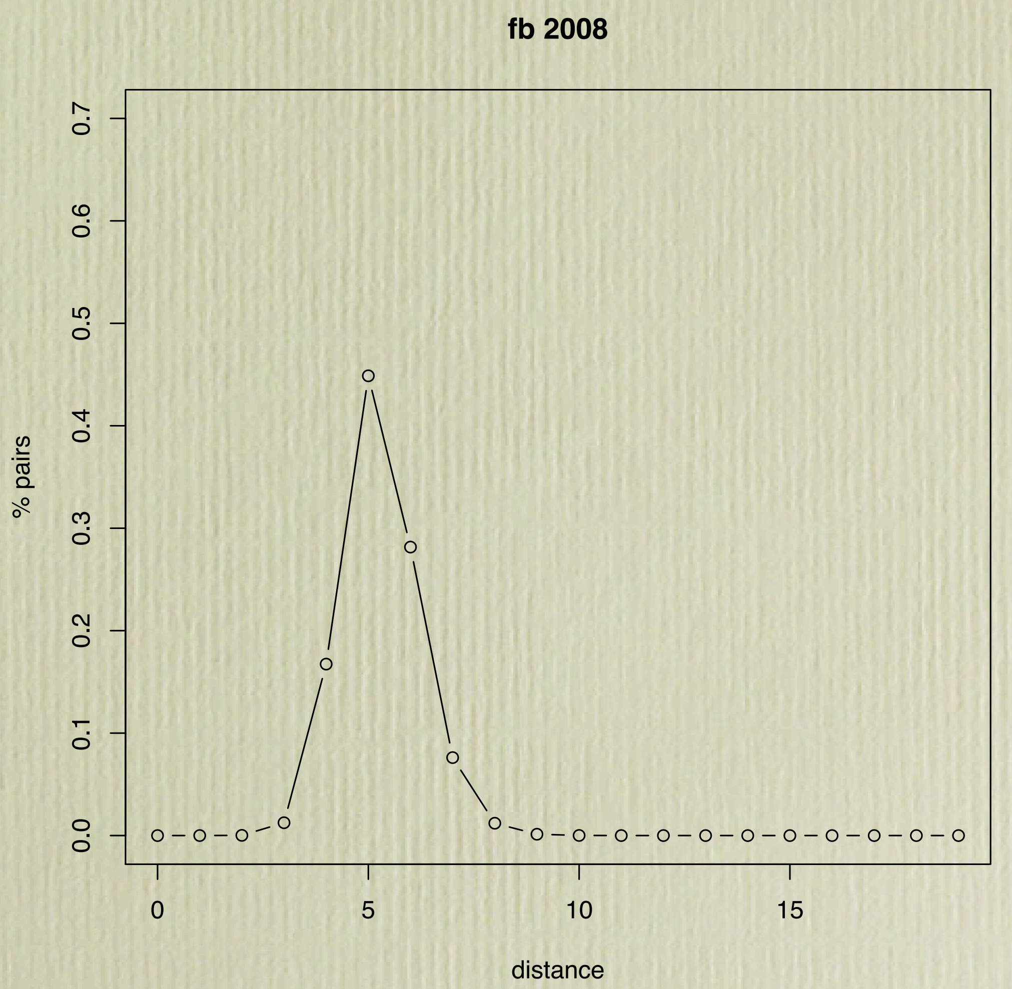
- We considered:
  - fb: the whole facebook graph
  - it / se: only Italian / Swedish users
  - it+se: only Italian & Swedish users
  - us: only US users
- Based on users' *current* geo-IP location

Distance distribution (fb)

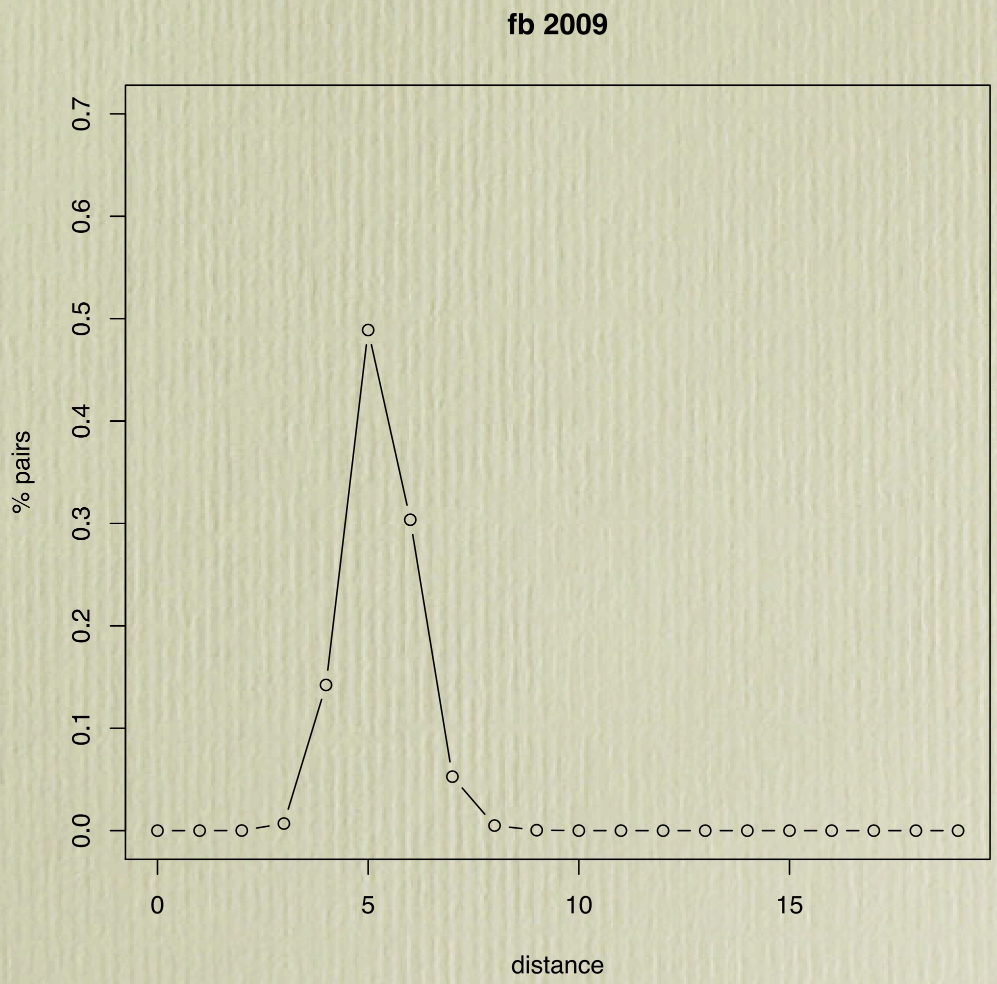
Distance distribution (fb)



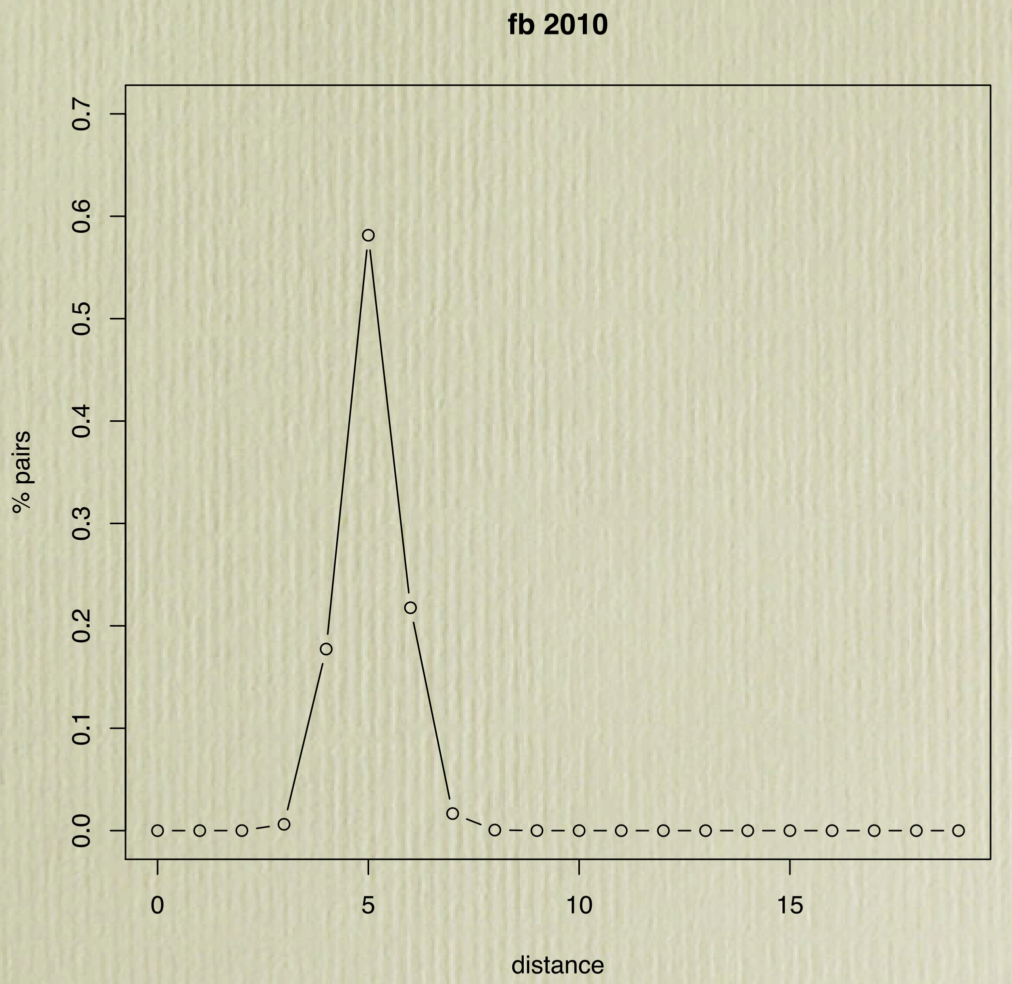
Distance distribution (fb)



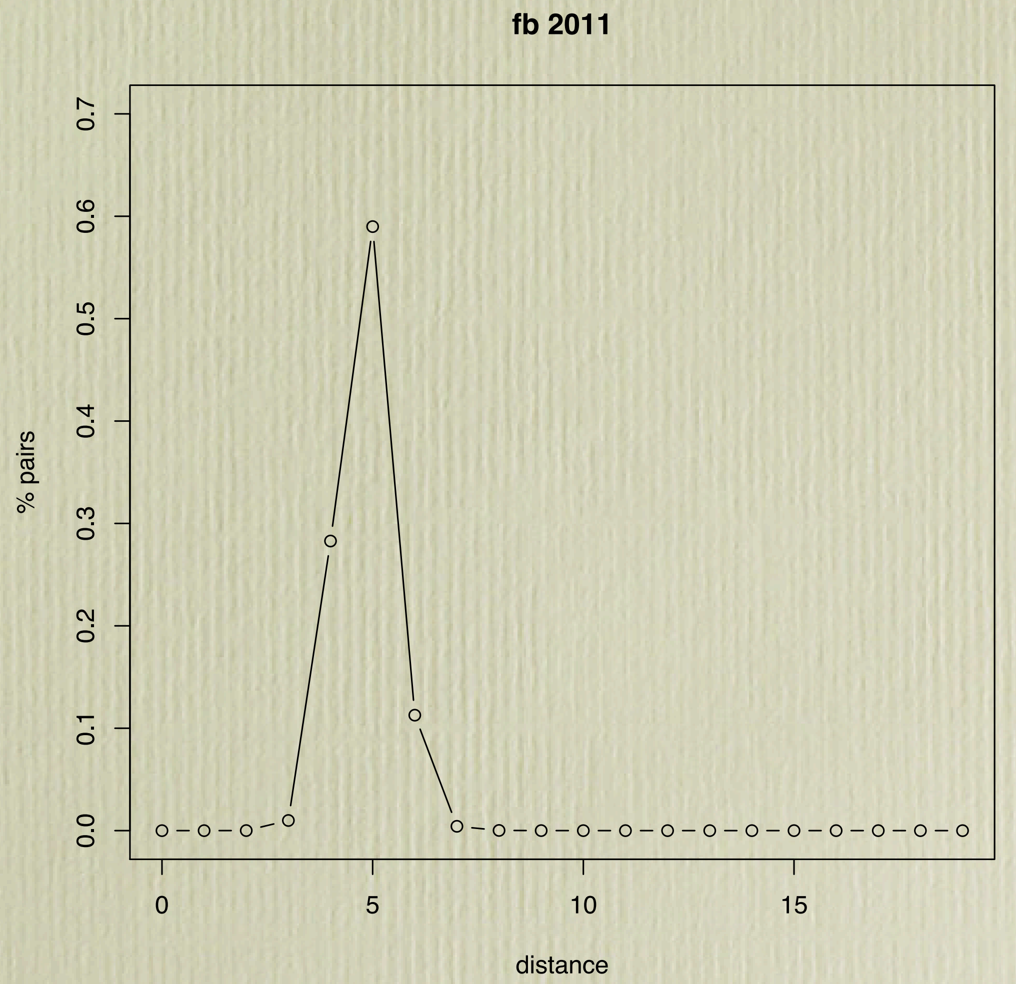
Distance distribution (fb)



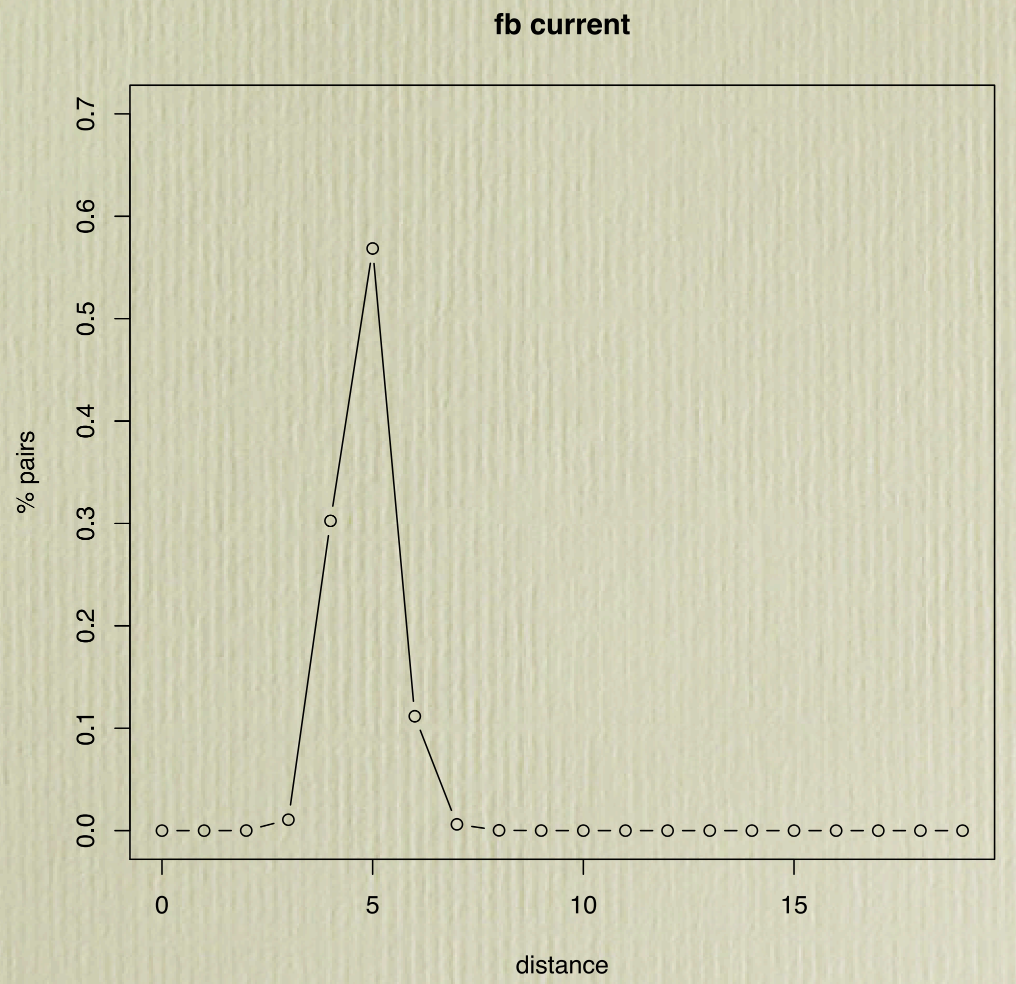
Distance distribution (fb)



Distance distribution (fb)

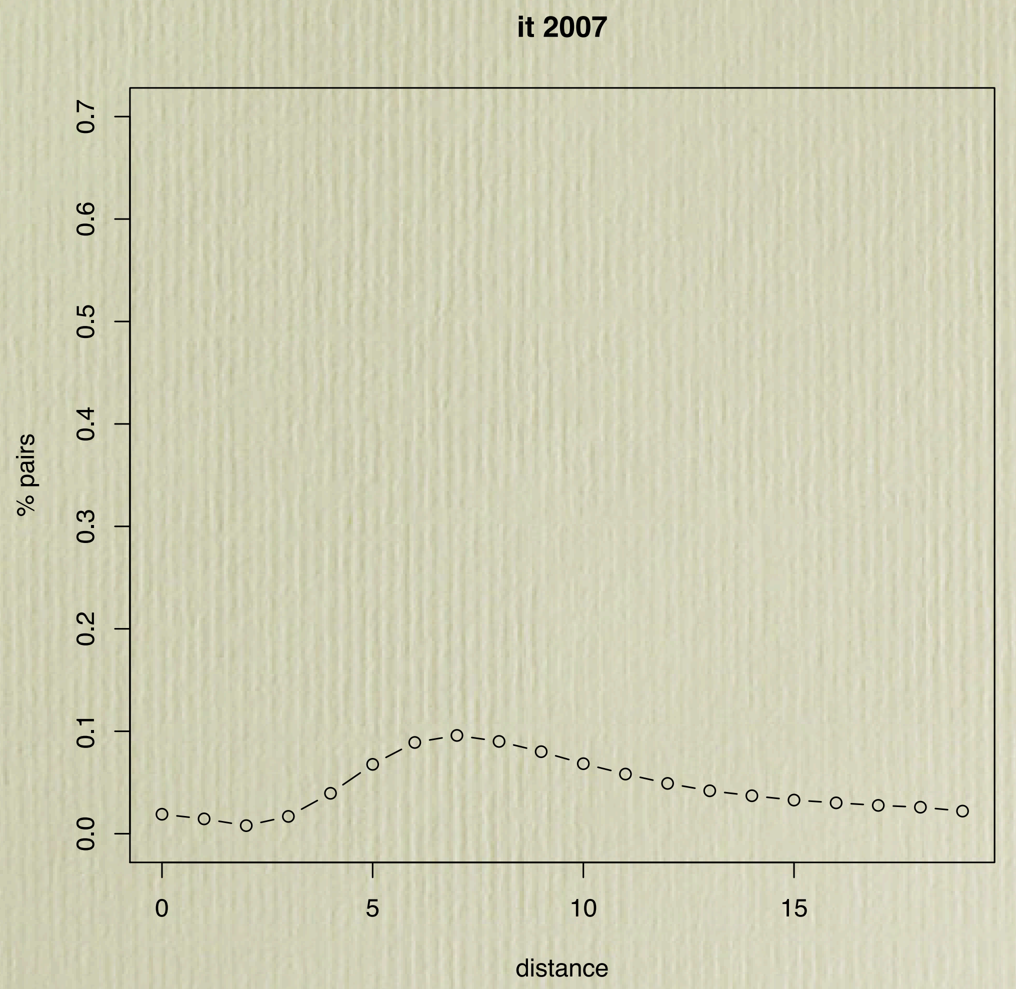


Distance distribution (fb)

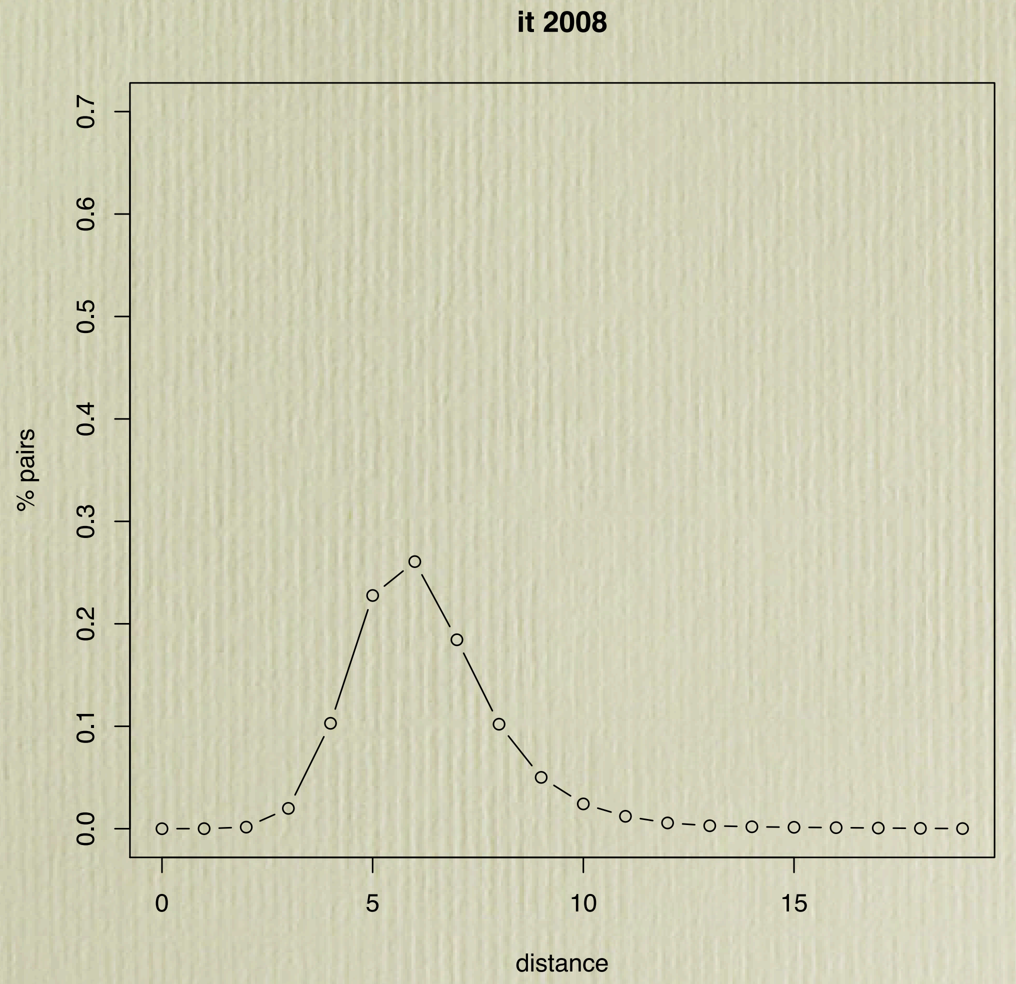


Distance distribution (it)

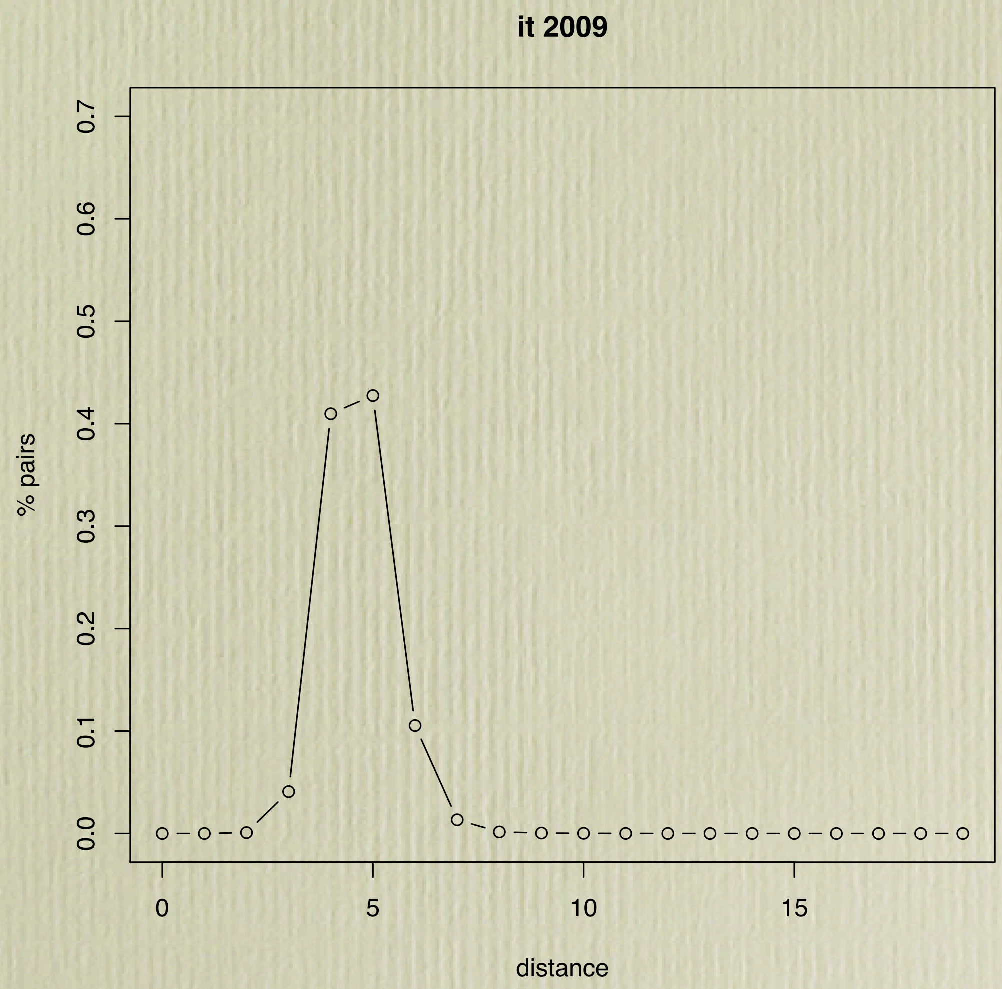
Distance distribution (it)



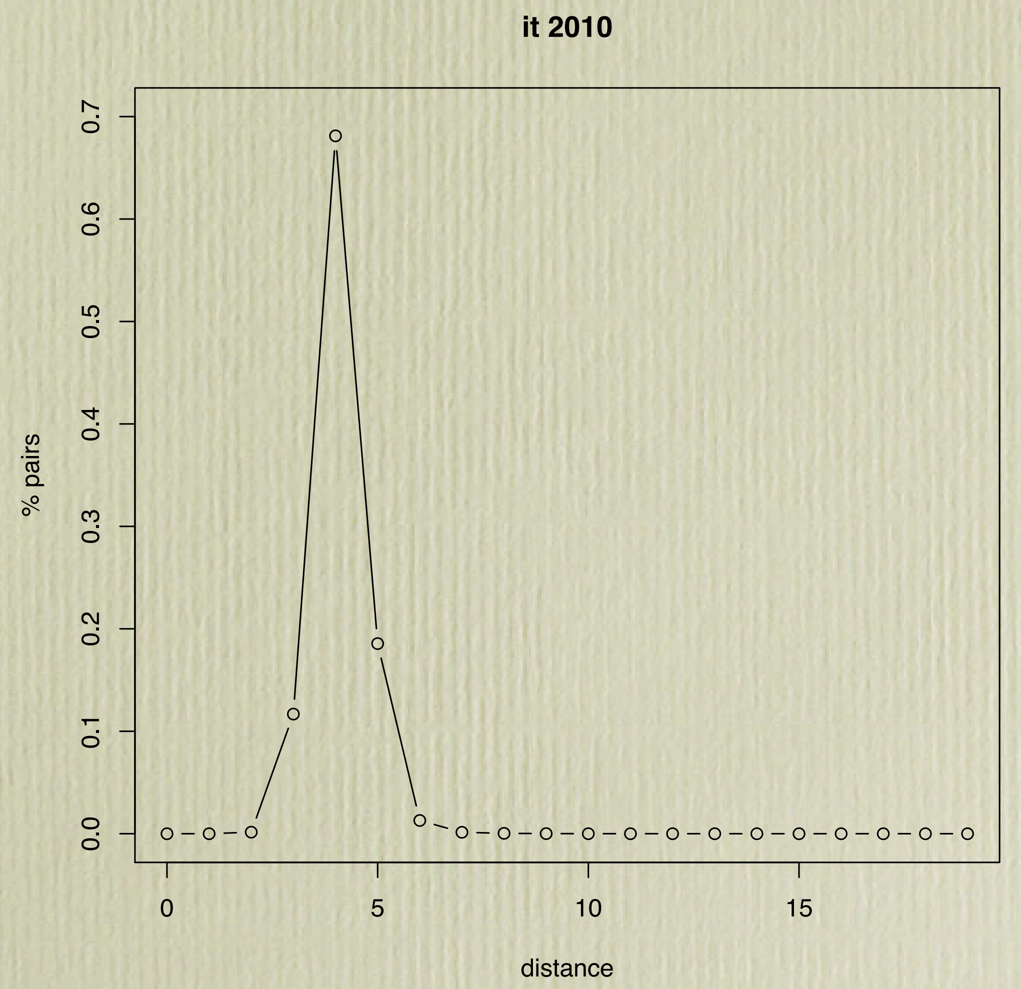
Distance distribution (it)



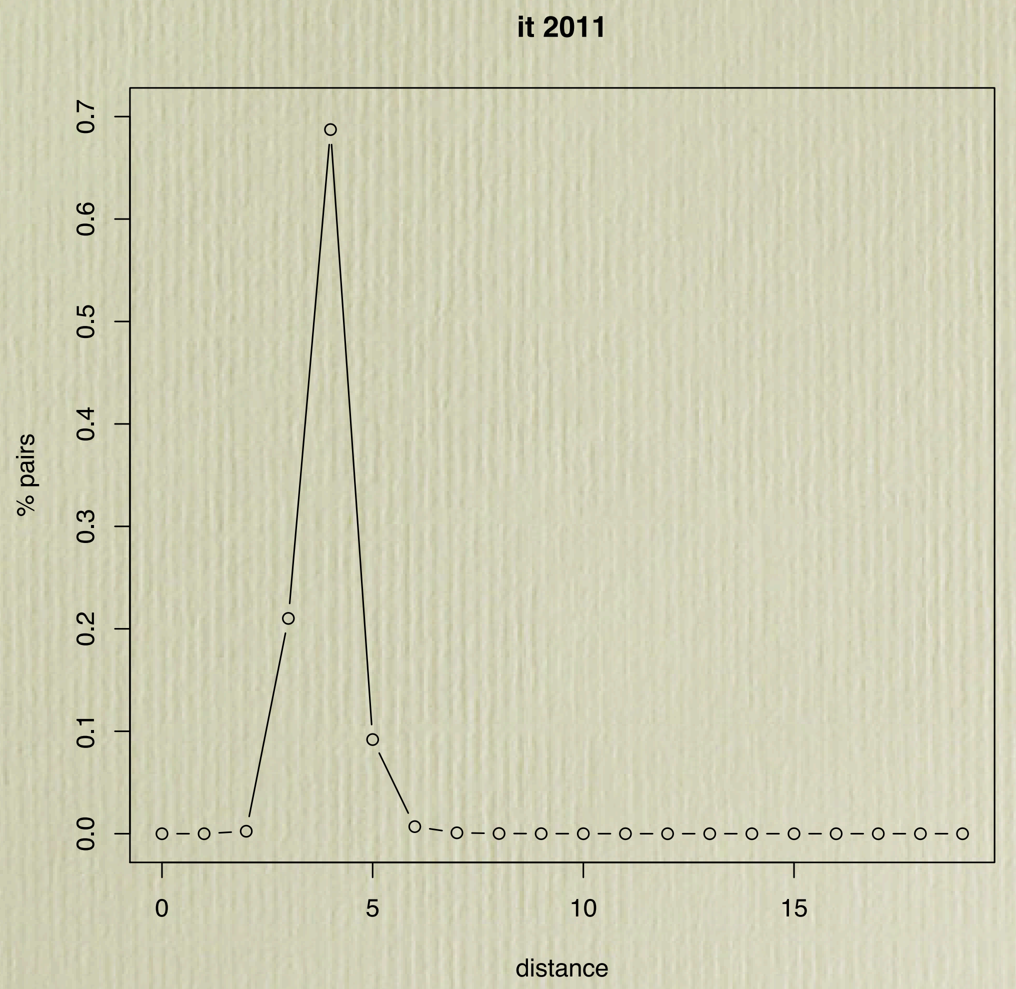
Distance distribution (it)



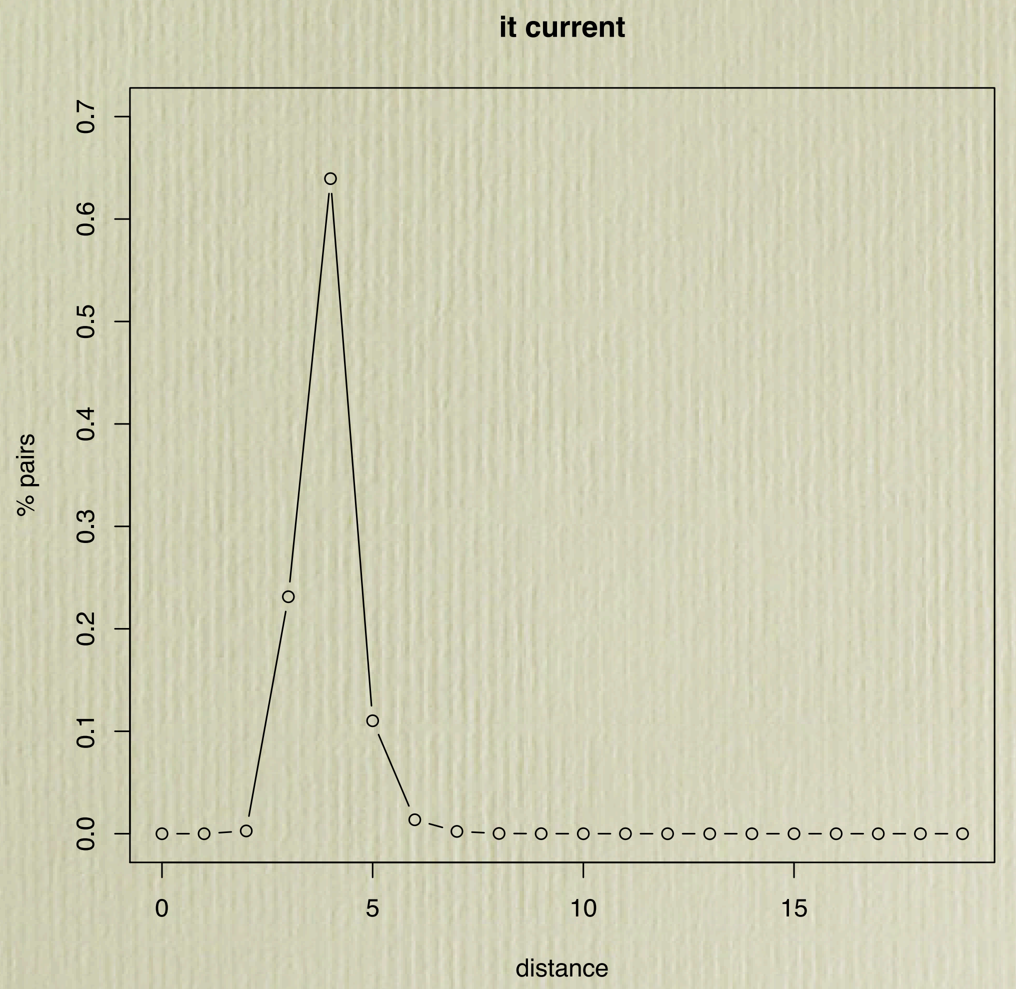
Distance distribution (it)



Distance distribution (it)

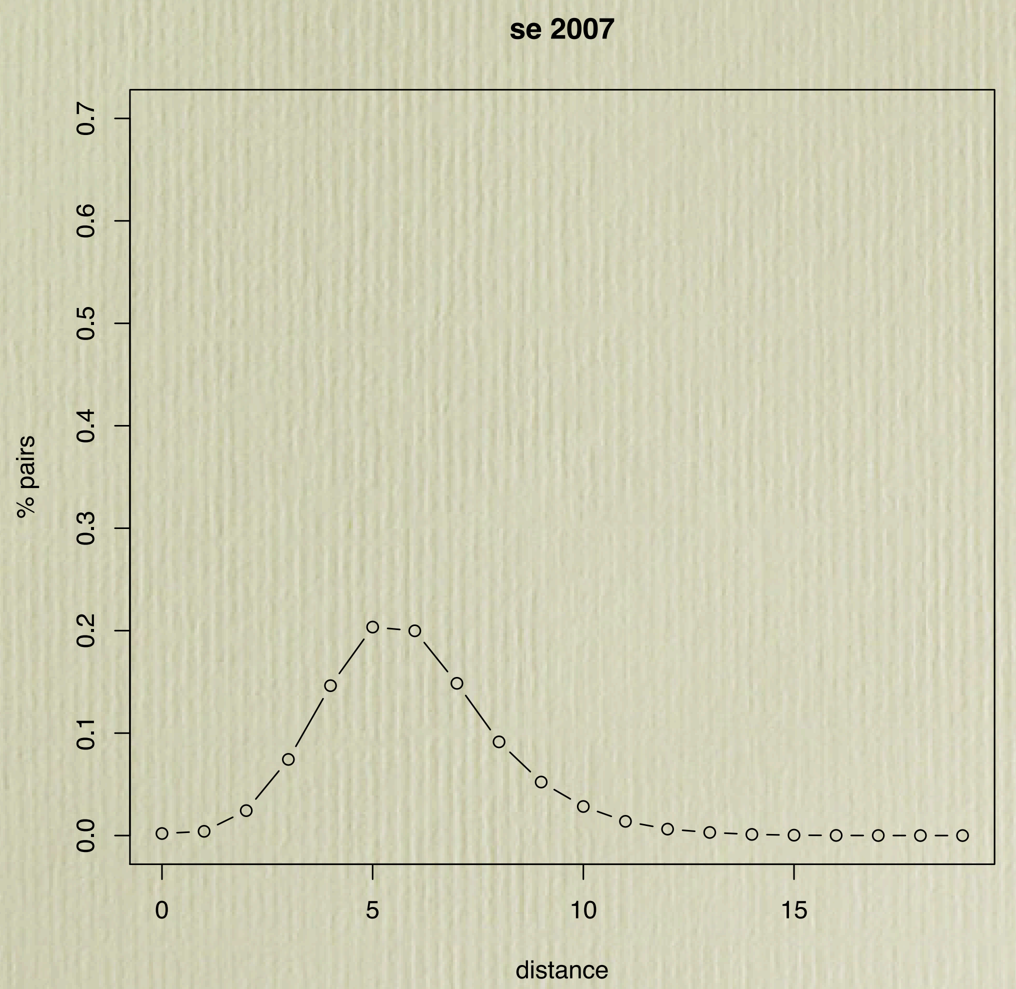


Distance distribution (it)

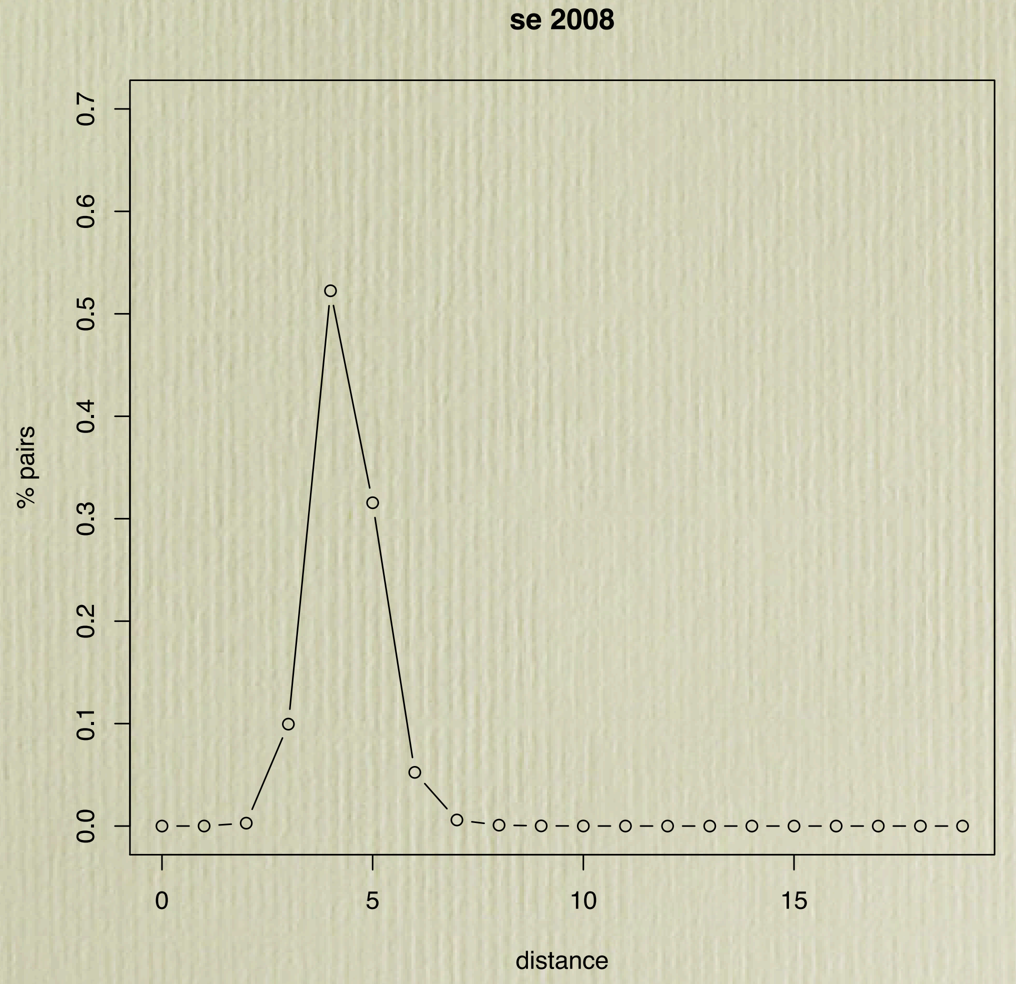


Distance distribution (se)

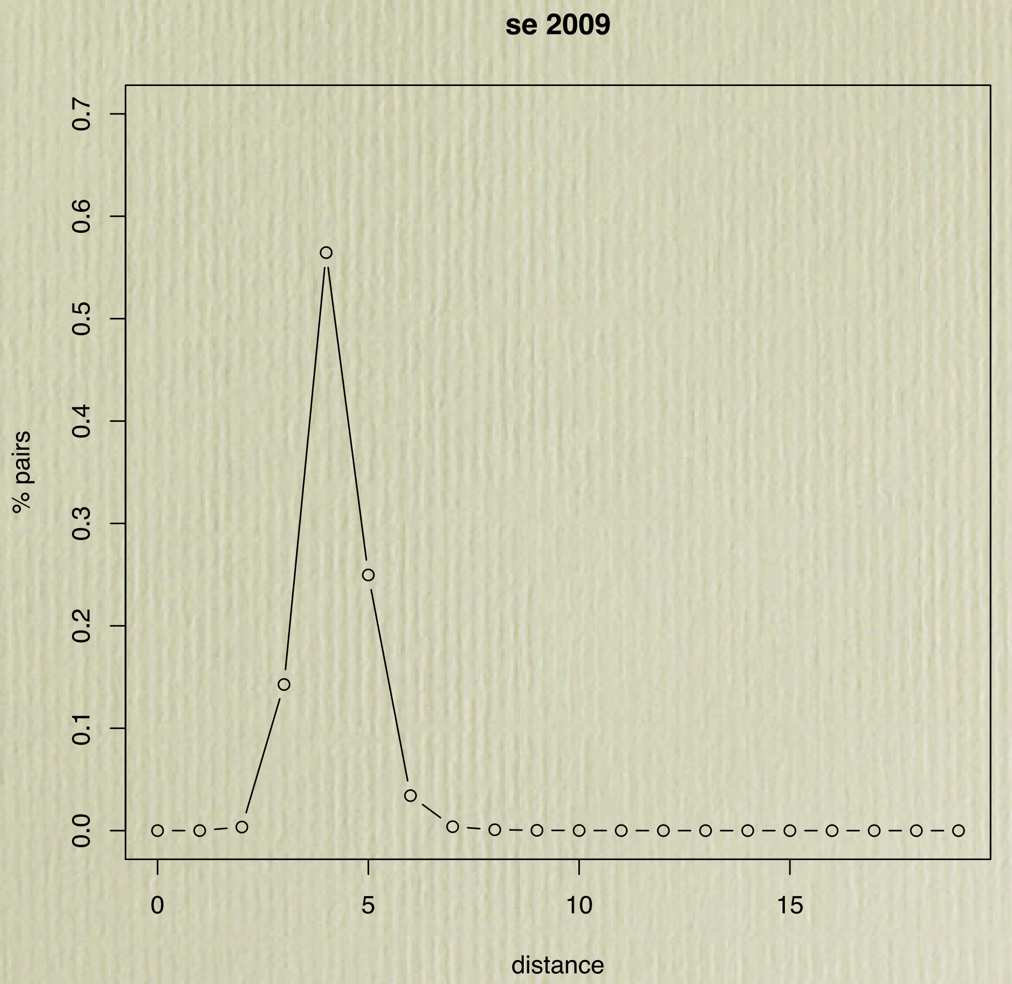
Distance distribution (se)



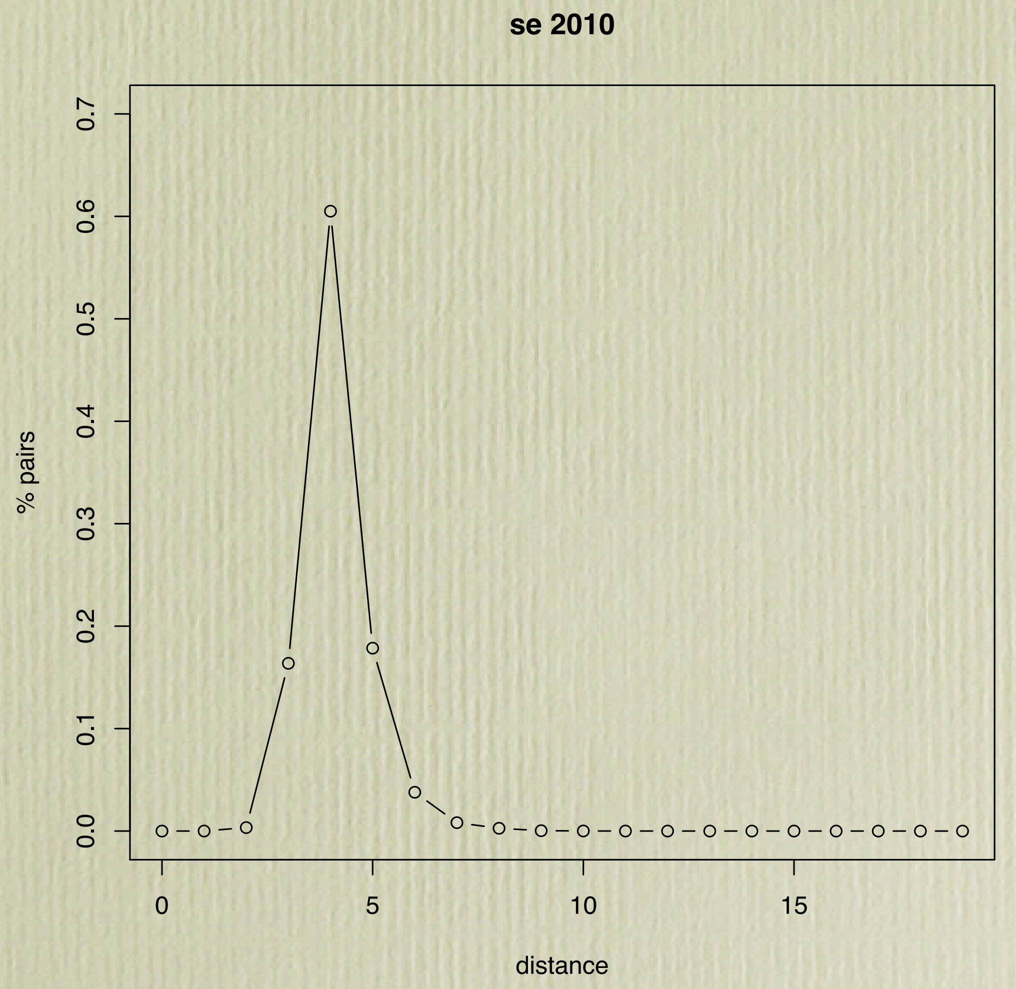
Distance distribution (se)



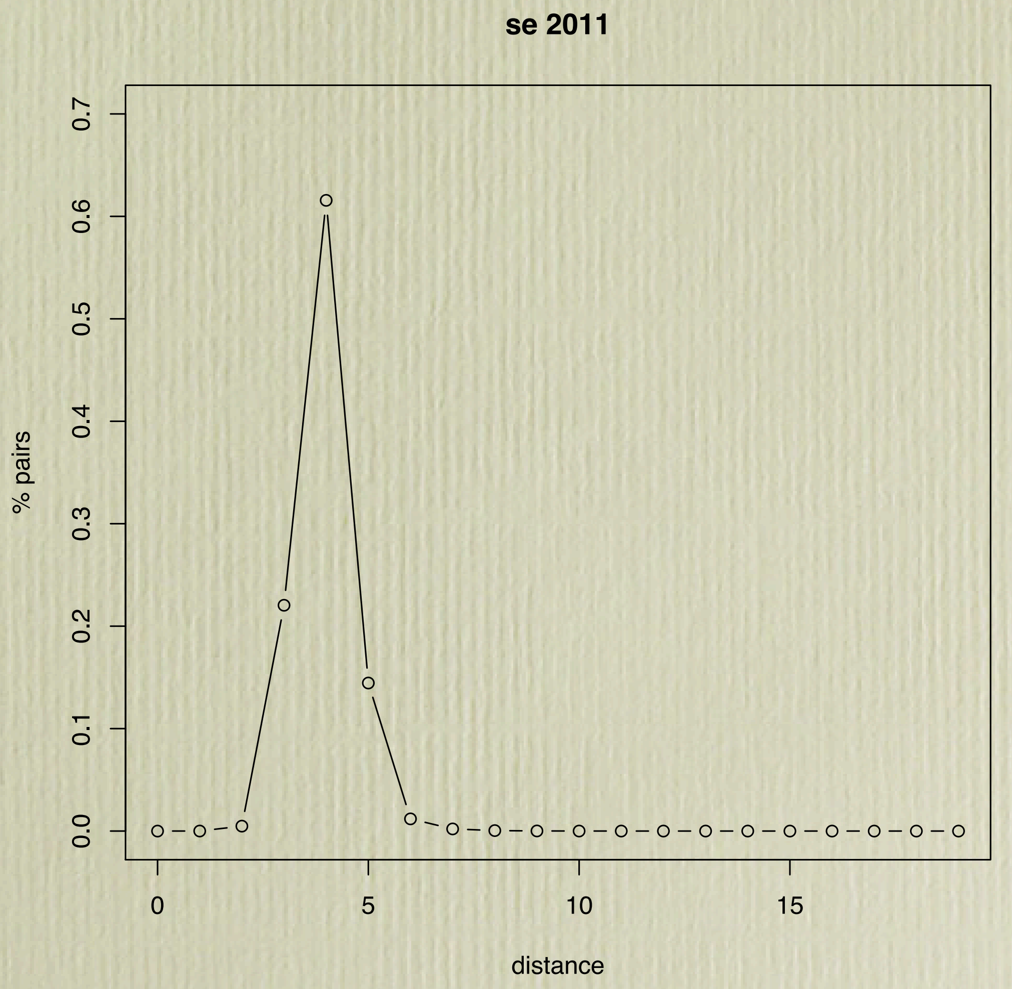
Distance distribution (se)



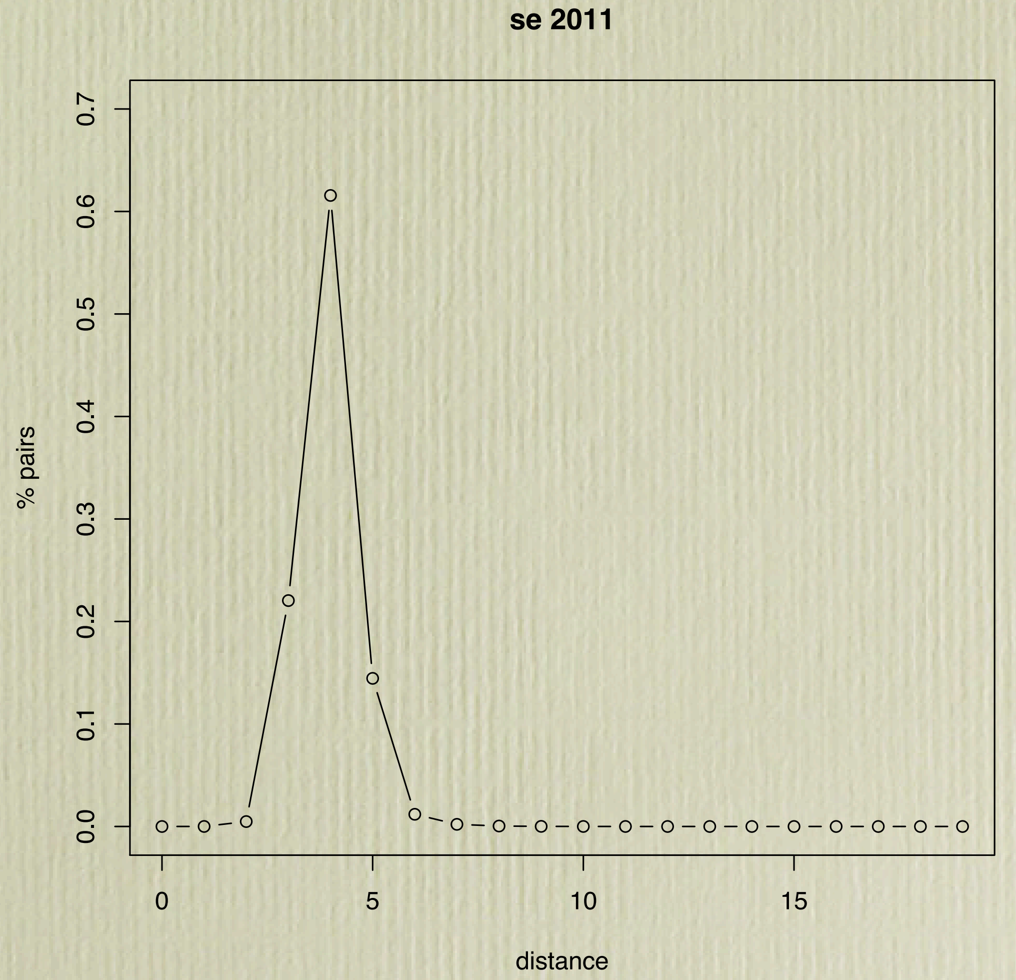
Distance distribution (se)



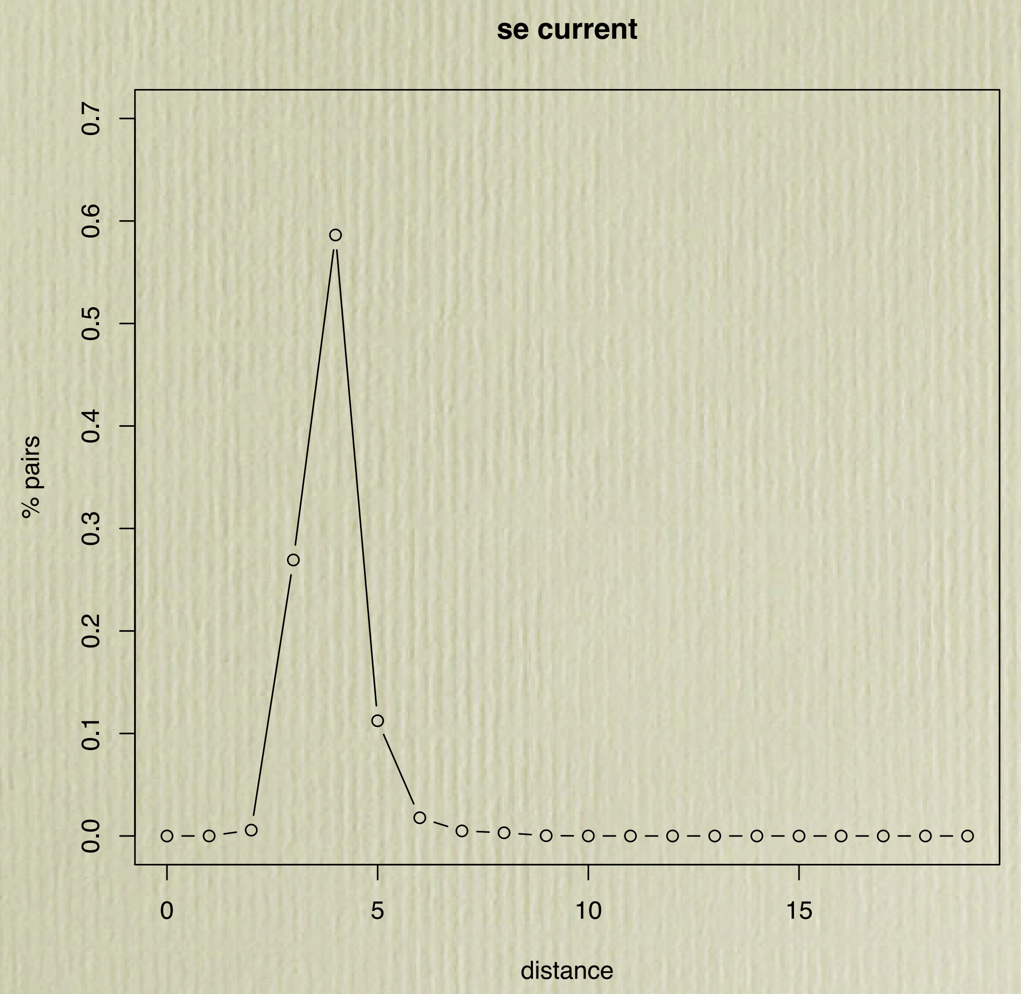
Distance distribution (se)



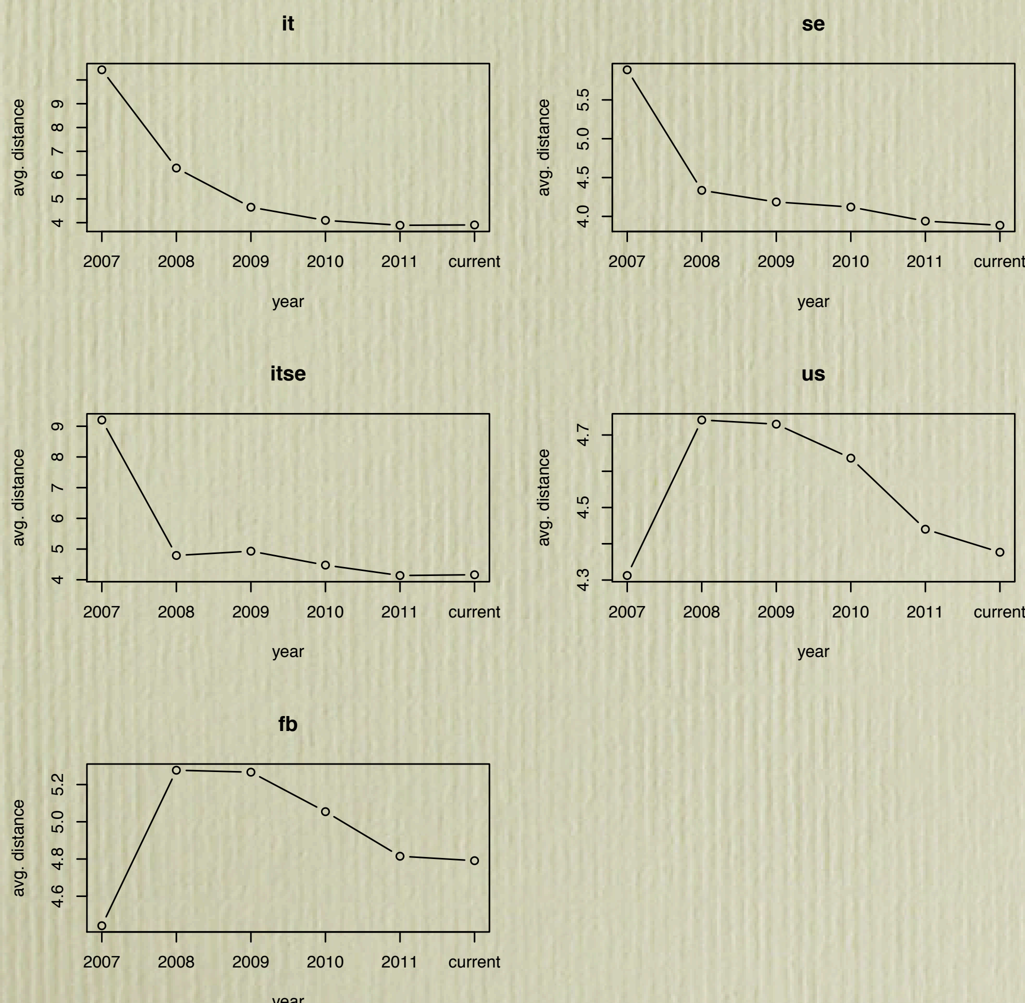
Distance distribution (se)



Distance distribution (se)

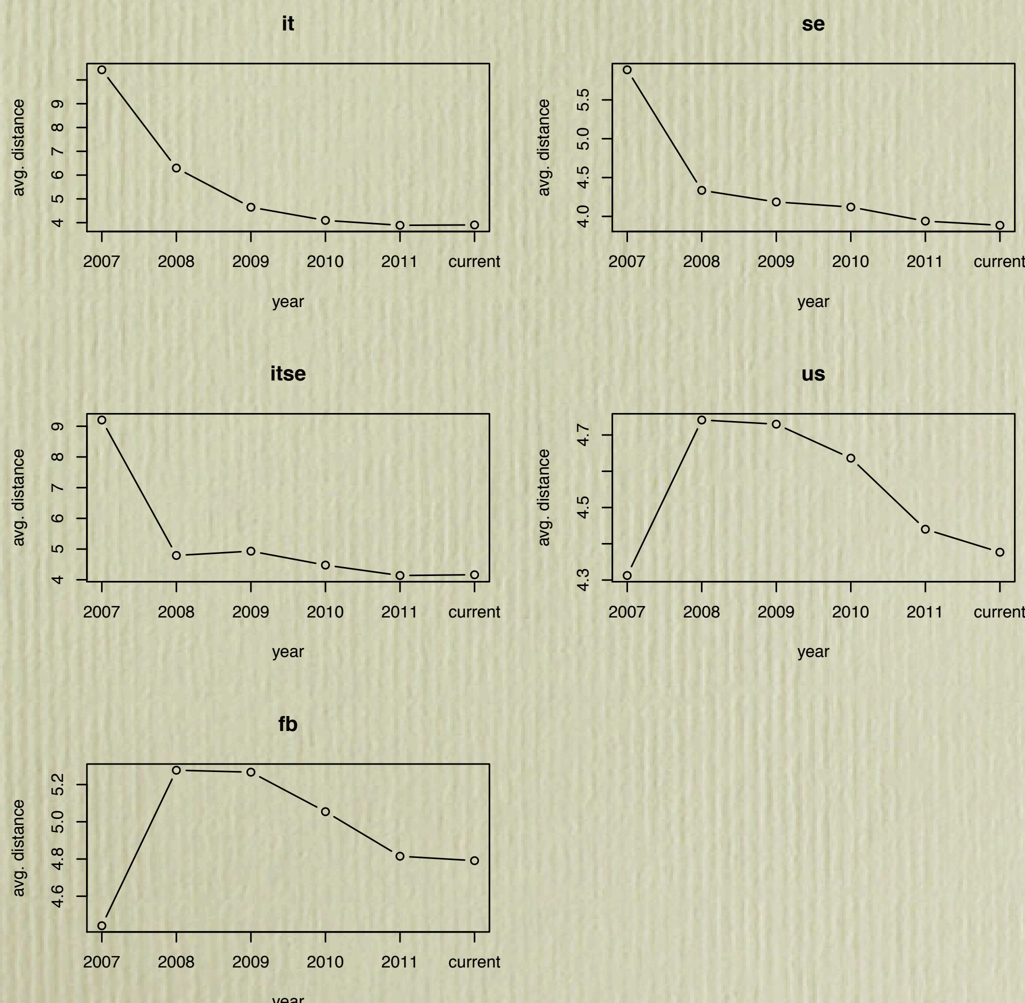


Average distance



	2008	*curr*
it	6.45	3.89
se	4.37	3.90
it+se	4.85	4.16
us	4.75	4.32
fb	5.28	4.74

Average distance



	2008	*curr*
it	6.45	3.89
se	4.37	3.90
it+se	4.85	4.16
us	4.75	4.32
fb	5.28	4.74

Average distance


HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times

Business Day

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SP

Separating You and Me? 4.74 Degrees

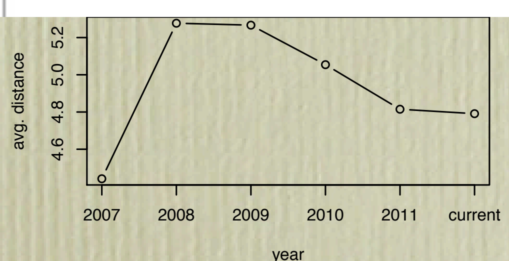
By JOHN MARKOFF and SOMINI SENGUPTA

Published: November 21, 2011

The world is even smaller than you thought.

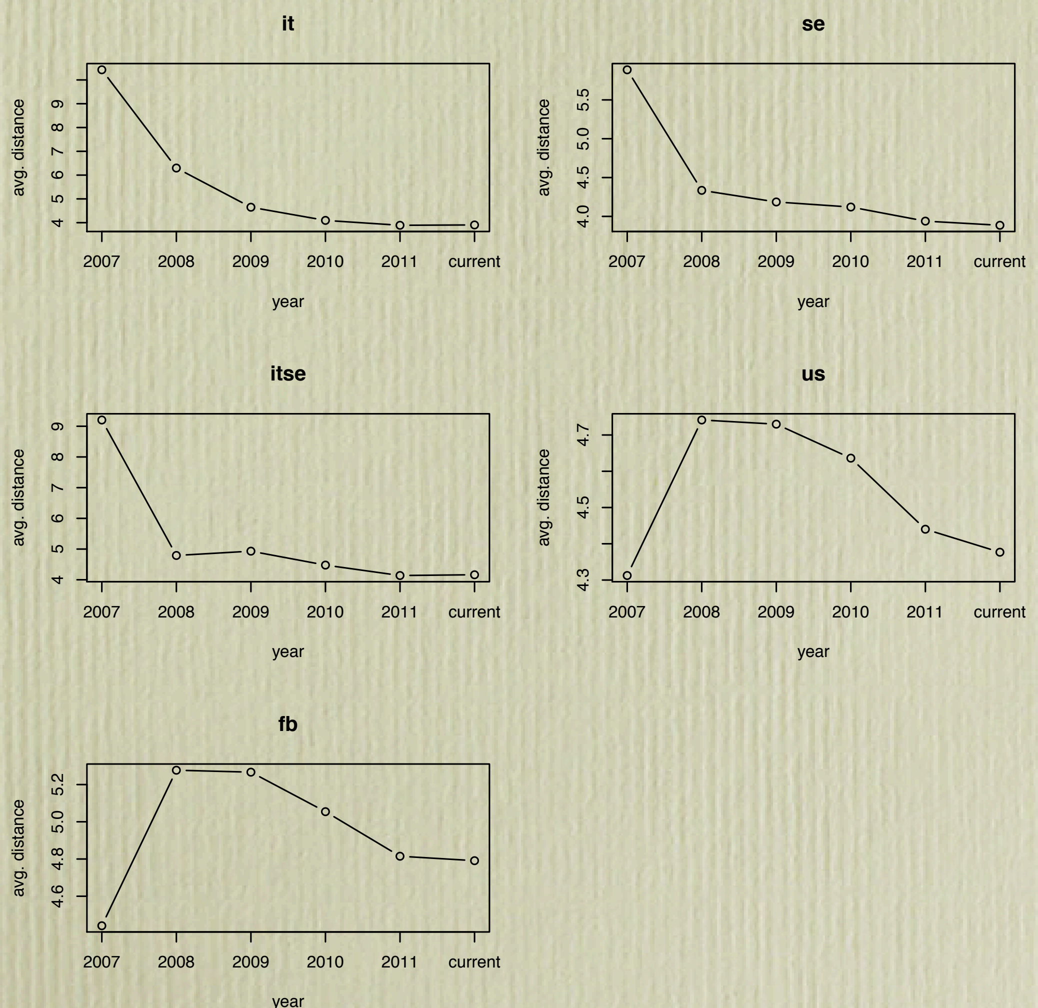
RECOMMEND

TWITTER



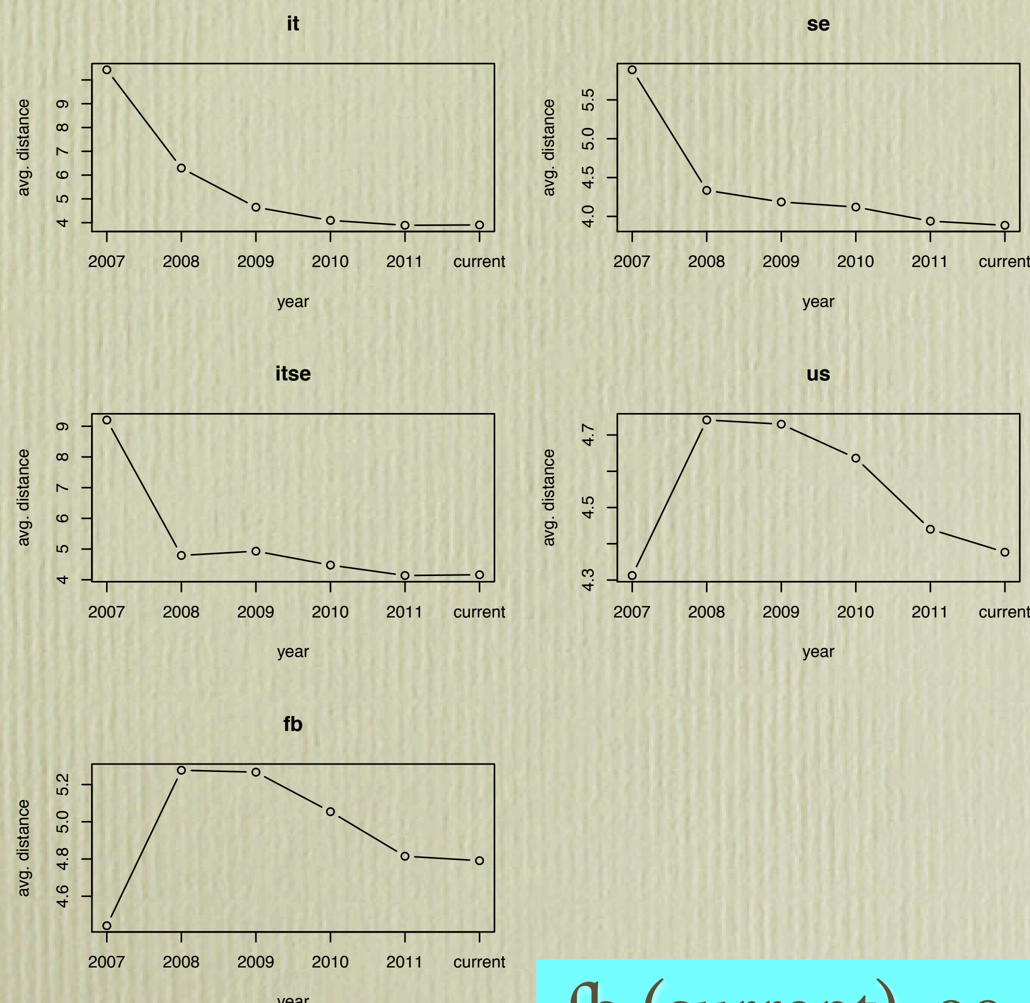
		…urr
		…89
		…90
		…16
		…32
fb	5.28	4.74

Average distance



	2008	curr
it	6.45	3.89
se	4.37	3.90
it+se	4.85	4.16
us	4.75	4.32
fb	5.28	4.74

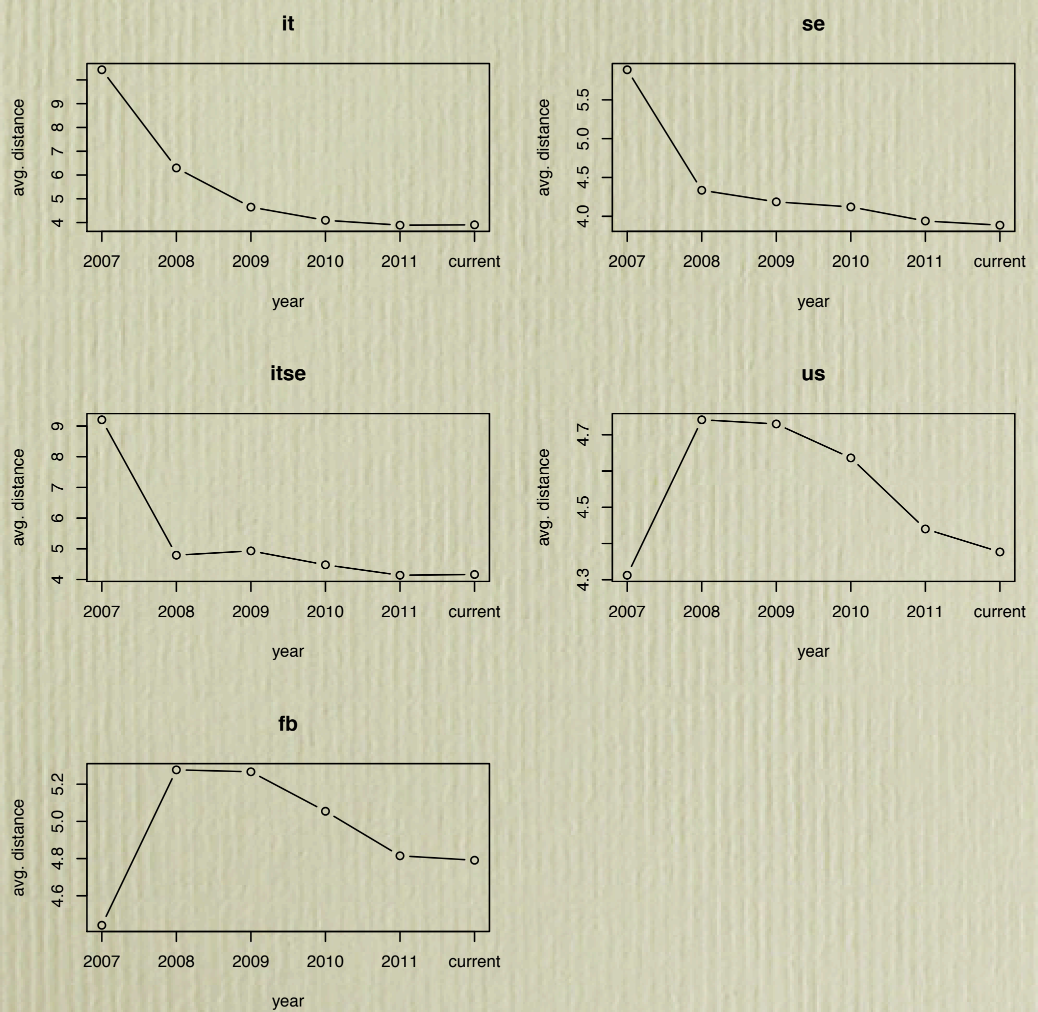
Average distance



	2008	*curr*
it	6.45	3.89
se	4.37	3.90
it+se	4.85	4.16
us	4.75	4.32
fb	5.28	4.74

fb (current): 99.9% nodes in the giant component!

Average distance



	2008	curr
it	6.45	3.89
se	4.37	3.90
it+se	4.85	4.16
us	4.75	4.32
fb	5.28	4.74

Average degree vs. density (fb)

	Avg. degree	*Density*
2009	88.7	$6.4 * 10^{-7}$
2010	113.0	$3.4 * 10^{-7}$
2011	169.0	$3.0 * 10^{-7}$
curr	190.4	$2.6 * 10^{-7}$

Diameter (max distance)

	2008	*curr*
it	≥28	=25
se	≥17	=23
it+se	≥24	=27
us	≥17	=30
fb	≥16	=41

Diameter (max distance)

Used the double-sweep lower bound/ iterative fringe upper bound technique (Crescenzi, Grossi, Habib, Lanzi & Marino, 2011)

	2008	*curr*
it	≥28	=25
se	≥17	=23
it+se	≥24	=27
us	≥17	=30
fb	≥16	=41

Another application: Spid

Another application: Spid

- We proposed to use the *spid* (*shortest-paths index of dispersion*), that is, the ratio between variance and mean of the distance distribution, as a network feature

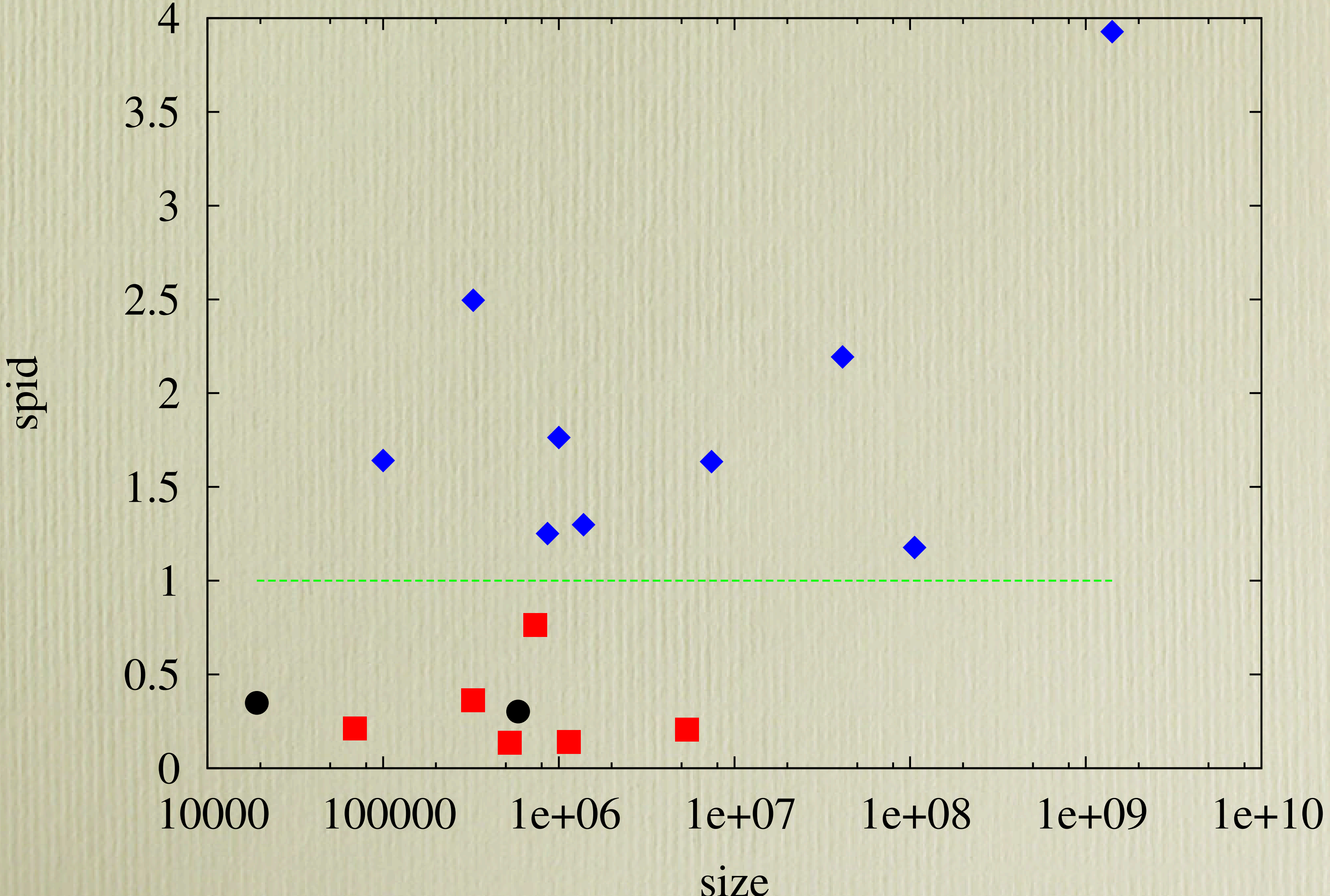
Another application: Spid

- We proposed to use the *spid* (*shortest-paths index of dispersion*), that is, the ratio between variance and mean of the distance distribution, as a network feature
- When the dispersion index is <1, the distribution is *underdispersed*; >1, is *overdispersed*

Another application: Spid

- We proposed to use the *spid* (*shortest-paths index of dispersion*), that is, the ratio between variance and mean of the distance distribution, as a network feature
- When the dispersion index is <1, the distribution is *underdispersed*; >1, is *overdispersed*
- Web graphs and social networks are **different** under this viewpoint!

Spid plot



4
3.5
3
2.5
2
1.5
1
0.5
0
spid
10000
100000
1e+06
1e+07
1e+08
1e+09
1e+10
size

Spid conjecture

Spid conjecture

- We conjectured that spid is able to tell social networks from web graphs

Spid conjecture

- We conjectured that spid is able to tell social networks from web graphs
- Average distance alone would not suffice: it is very changeable and depends on the scale

Spid conjecture

- We conjectured that spid is able to tell social networks from web graphs
- Average distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1

Spid conjecture

- We conjectured that spid is able to tell social networks from web graphs
- Average distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1
- What is Facebook spid?

Spid conjecture

- We conjectured that spid is able to tell social networks from web graphs
- Average distance alone would not suffice: it is very changeable and depends on the scale
- Spid, instead, seems to have a clear cutpoint at 1
- What is Facebook spid?

[Answer: 0.09]

Do it yourself!

Do it yourself!

- HyperANF is available within the WebGraph framework

Do it yourself!

- HyperANF is available within the WebGraph framework
- Download it from http://webgraph.di.unimi.it/

Do it yourself!

- HyperANF is available within the WebGraph framework
- Download it from http://webgraph.di.unimi.it/
- Or google for “web graph”

Do it yourself!

- HyperANF is available within the WebGraph framework
- Download it from http://webgraph.di.unimi.it/
- Or google for “web graph”
- Lots of social networks ready to download at http://law.di.unimi.it/

Do it yourself!

- HyperANF is available within the WebGraph framework
- Download it from http://webgraph.di.unimi.it/
- Or google for “web graph”
- Lots of social networks ready to download at http://law.di.unimi.it/
- Distributions analysed in this paper available, too

Questions

Not all pairs are connected: how can the average distance be even finite?

Interesting question

Interesting question

- Here by *average distance* we mean *average over all reachable pairs*

Interesting question

- Here by *average distance* we mean *average over all reachable pairs*
- The number of reachable pairs is a sort of *confidence*: in our case, it is 99.9%

Interesting question

- Here by *average distance* we mean *average over all reachable pairs*
- The number of reachable pairs is a sort of *confidence*: in our case, it is 99.9%
- The latter is an important datum

Interesting question

- Here by *average distance* we mean *average over all reachable pairs*
- The number of reachable pairs is a sort of *confidence*: in our case, it is 99.9%
- The latter is an important datum
  - after all, a disconnected graph of 1 million nodes has average distance 0, but with 0.00001% confidence

What about Milgram?

What about Milgram?

- Very difficult even to state this in Milgram's setting

What about Milgram?

- Very difficult even to state this in Milgram's setting
- If we assume that all uncompleted chains correspond to unreachable pairs, the confidence of his measure was 22% (or 29%, if we consider only chains that at least started)

An alternative, anybody?

An alternative, anybody?

- Alternatively, one can consider the *harmonic diameter* (the harmonic mean of *all* distances):

An alternative, anybody?

- Alternatively, one can consider the *harmonic diameter* (the harmonic mean of *all* distances):

$$\frac{n(n-1)}{\sum_{x \neq y} \frac{1}{d(x,y)}}$$

An alternative, anybody?

- Alternatively, one can consider the *harmonic diameter* (the harmonic mean of *all* distances):

$$\frac{n(n-1)}{\sum_{x \neq y} \frac{1}{d(x,y)}}$$

- where the summation is extended to all pairs of distinct nodes, and the reciprocal of infinity is assumed to be 0 (Marchiori & Latora, 2000)

An alternative, anybody?

- Alternatively, one can consider the *harmonic diameter* (the harmonic mean of *all* distances):

$$\frac{n(n-1)}{\sum_{x \neq y} \frac{1}{d(x,y)}}$$

- where the summation is extended to all pairs of distinct nodes, and the reciprocal of infinity is assumed to be 0 (Marchiori & Latora, 2000)
- Milgrams's harmonic diameter for the random sample is 26.68!

Harmonic diameter

Harmonic diameter

	2008	*curr*
it	23.7	3.68
se	4.37	3.69
it+se	6.4	3.90
us	4.61	4.45
fb	5.74	4.59

Harmonic diameter

	2008	*curr*
it	23.7	3.68
se	4.37	3.69
it+se	6.4	3.90
us	4.61	4.45
fb	5.74	4.59

Compare with average distance

	2008	*curr*
it	6.58	3.90
se	4.33	3.89
it+se	4.9	4.16
us	4.74	4.32
fb	5.28	4.74

Harmonic diameter

	2008	*curr*
it	23.7	3.68
se	4.37	3.69
it+se	6.4	3.90
us	4.61	4.45
fb	5.74	4.59

Compare with average distance

An alternative: use median (similar outcomes)

	2008	*curr*
it	6.58	3.90
se	4.33	3.89
it+se	4.9	4.16
us	4.74	4.32
fb	5.28	4.74

The sample is biased, and anyway it just represents 10% of humanity!

Uniform?

Uniform?

- Facebook is not a uniform sample (if anything, because of digital divide)

Uniform?

- Facebook is not a uniform sample (if anything, because of digital divide)
- But 96 people from Nebraska are not a random sample of humanity, either

Friendship?

Friendship?

- Is the notion of friendship in Facebook an approximation of the notion of friendship in real life?

Friendship?

- Is the notion of friendship in Facebook an approximation of the notion of friendship in real life?
- The notion of friendship used by Milgram (*first-name acquaintance*) may be even weaker!

You measured the average distance, but degrees of separation are algorithmic

Or, is it?

Or, is it?

- The point is the distinction between *routing* (a.k.a. functional degree of separation) and *distance*

Or, is it?

- The point is the distinction between *routing* (a.k.a. functional degree of separation) and *distance*
- The interest in efficient routing lies more in the eye of the beholder (e.g., the computer scientist) than in Milgram's

Or, is it?

- The point is the distinction between *routing* (a.k.a. functional degree of separation) and *distance*
- The interest in efficient routing lies more in the eye of the beholder (e.g., the computer scientist) than in Milgram's
- Reading carefully Travers and Milgram's papers, it is clear that they had distance and not routing in mind:

Or, is it?

- The point is the distinction between *routing* (a.k.a. functional degree of separation) and *distance*
- The interest in efficient routing lies more in the eye of the beholder (e.g., the computer scientist) than in Milgram's
- Reading carefully Travers and Milgram's papers, it is clear that they had distance and not routing in mind:

given two individuals selected randomly
from the population, what is the
probability that the *minimum* number of
intermediaries required to
link them is 0, 1, 2, ...

Just add a few links here and there and we'll all be at one degree of separation

How true is this statement?

How true is this statement?

- Suppose that we consider *any* network with the same number of edges m, the same maximum degree D and the same number of reachable pairs of nodes r

How true is this statement?

- Suppose that we consider *any* network with the same number of edges m, the same maximum degree D and the same number of reachable pairs of nodes r
- How small can the average distance be?

How true is this statement?

- Suppose that we consider *any* network with the same number of edges m, the same maximum degree D and the same number of reachable pairs of nodes r
- How small can the average distance be?
- Exactly m pairs at distance 1, at most mD pairs at distance 2, and all other pairs at distance 3 or greater...

So...

So...

- With the Facebook data ($\underset{\sim}{m}$=69E9, r=5E17, D=5000, n=721E6) we obtain that the average distance cannot be smaller than 2.999

So...

- With the Facebook data (*m*=69E9, r=5E17, D=5000, n=721E6) we obtain that the average distance cannot be smaller than 2.999
- In other words, only increasing the degree and/or increasing the density we could go below 3...

So...

- With the Facebook data ($m_{\sim}$=69E9, r=5E17, D=5000, n=721E6) we obtain that the average distance cannot be smaller than 2.999
- In other words, only increasing the degree and/or increasing the density we could go below 3...
- Our measured value (4.74) is not so far from this lower bound

Moreover...

Moreover...

- We can refine this analysis to a bound depending on the *degree sequence* (Boldi & Vigna, 2012)

Moreover...

- We can refine this analysis to a bound depending on the *degree sequence* (Boldi & Vigna, 2012)
- Plugging in the Facebook degree sequence we obtain a lower bound of 3.6

Moreover...

- We can refine this analysis to a bound depending on the *degree sequence* (Boldi & Vigna, 2012)
- Plugging in the Facebook degree sequence we obtain a lower bound of 3.6
- This means that no graph with the same degree distribution can go below this lower bound

Moreover...

- We can refine this analysis to a bound depending on the *degree sequence* (Boldi & Vigna, 2012)
- Plugging in the Facebook degree sequence we obtain a lower bound of 3.6
- This means that no graph with the same degree distribution can go below this lower bound
- Again, notice the small gap with 4.74...

More questions?