Gossip Protocols

Márk Jelasity

Hungarian Academy of Sciences and
University of Szeged, Hungary

Introduction

- Gossip-like phenomena are commonplace
  - human gossip
  - epidemics (virus spreading, etc)
  - computer epidemics (malicious agents: worms, viruses)
  - phenomena such as forest fires, branching processes and diffusion are all similar mathematically
- extremely simple locally, powerful and robust globally
- In computer science, epidemics are relevant
  - for security (against worms and viruses)
  - for designing useful protocols (we look at this here)

Outline

- Information dissemination
  - Seminal work by Demers *et al* (1987), that first coined the term gossip and epidemic protocols
  - A few words on random and complex networks
- Generalizations of gossip protocols for
  - peer sampling
  - topology maintenance
  - data aggregation
  - modular architectures
- Problems, directions

Epidemic Database Updates

- Problem
  - Xerox corporate Internet, replicated databases
  - Each database has a set of keys that have values (along with a time stamp)
  - Goal: all databases are the same, in the face of key updates, removals and additions
  - Updates are made locally and have to be replicated at all sites (300 sites)
- Solution in 1986: emailing updates
  - problems with detecting and correcting errors (done by hand!)
  - bottleneck with the originating (updated) site
  - not scalable (slow if very large number of nodes)
  - (message complexity quite good though!)

Gossip to the rescue

- Main components are replaced by gossip
  - update spreading: rumor mongering (no bottleneck)
  - error correction: anti-entropy gossip (reliable)
- anti-entropy
  - uses “simple epidemics” with two states: infective and susceptible (a.k.a. SI model)
  - guarantees perfect dissemination
- rumor mongering
  - uses “complex epidemics” with an additional state: removed (a.k.a. SIR model)
  - certain (quite small) probability of error

Some Properties of the SI model

- the push model
  - N nodes communicate in rounds (cycles)
  - in each cycle, a node that has the update (infected) sends it to a random other node, that becomes infected too
- In anti-entropy
  - nodes send the (hash of) the entire database (not only a single update)
    - **as a side effect, all new updates are spread according to the SI model**
  - receiving nodes update their own database via merging the unseen updates

Mean-field model of push SI

- Let p_i be the proportion of *not* infected nodes in cycle i
- $1-p_0=1/N$
- Pittel (1987) shows that the model below is quite accurate for predicting p_i

$$E(p_{i+1}) = p_i\left(1-\frac{1}{N}\right)^{N(1-p_i)} \approx p_i e^{-(1-p_i)}$$

Speed and cost of push SI

- Let S_N be the first cycle where p_i=0
- Pittel (1987) shows that in probability

$$S_N = \log(N) + \ln(N) + O(1)$$

- This is quite fast...
- But the number of overall messages sent is

$$O(N \log N)$$

Pull and push-pull SI

- With pull, we have

$$E\left(p_{i+1}\right)=p_i^2$$

- This is *very* fast when p_i is small (end phase)...
- Karp *et al* (2000) show that the number of overall messages sent with push-pull is

$$O\left(N\log\log N\right)$$

- But termination is trickier when no updates are available (for anti-entropy does not matter)

SIR for spreading single updates

- For anti-entropy, use a pull or push-pull SI modell
- For the spreading of updates, the termination problem needs to be addressed: rumor mongering with SIR model
- Push approach
  – when a rumor (update) becomes “cold”, stop pushing
- Pull approach
  – same as push, only stop offering update when pulled when it becomes cold

Rumor mongering with push

$$\frac{ds}{dt} = -si$$

$$\frac{di}{dt} = si - \frac{1}{k}(1-s)i$$

$$\rightarrow s = e^{-(k+1)(1-s)}$$

- Stop spreading info with probability 1/k if unsuccessful infection attempt (become removed)
- s: susceptible, i: infective, r: removed
- Eg if k=1, 20% miss the gossip, if k=2, 6% miss it
  - In general, with push, the prob to miss the update is approx e^{-m} (m is the overall messages)

Some other rumor mongering algorithms

- Removal algorithms
  - Counter: removed after exactly k unsuccessful attempts
  - Random: removed with pr. 1/k after each unsuccessful attempt
- Blind: removal algorithm is run in each cycle irrespective of contacted node
- Feedback: removal algorithm runs if contacted node was not susceptible

Some empirical results (1000 nodes)

Feedback+ Counter+ push

Counter k	Residue s	Traffic m	Convergence t_{ave}	t_{last}
1	0.176	1.74	11.0	16.8
2	0.037	3.30	12.1	16.9
3	0.011	4.53	12.5	17.4
4	0.0036	5.64	12.7	17.5
5	0.0012	6.68	12.8	17.7

Blind+ Random+ push

Counter k	Residue s	Traffic m	Convergence t_{ave}	t_{last}
1	0.960	0.04	19	38
2	0.205	1.59	17	33
3	0.060	2.82	15	32
4	0.021	3.91	14.1	32
5	0.008	4.95	13.8	32

Feedback+ Counter+ pull

Counter k	Residue s	Traffic m	Convergence t_{ave}	t_{last}
1	3.1×10^{-2}	2.70	9.97	17.63
2	5.8×10^{-4}	4.49	10.07	15.39
3	4.0×10^{-6}	6.09	10.08	14.00

Summary

- Spreading updates: email or rumor mongering?
  - both focus on a single update that eventually stops spreading
  - both have a certain probability of error
  - gossip has no bottleneck but it generates more messages in total
  - gossip is much cheaper to restart (dies out quickly if update is already known by most nodes)
- Anti-entropy gossip
  - very expensive because looks at entire database
  - but fixes any distribution errors with prob. 1

Combining anti-entropy and rumor mongering

- Rumor mongering is used to spread updates
- Anti-entropy is run infrequently to make sure all updates are spread with pr. 1
- When anti-entropy finds an undelivered update: redistribution
  - Redistribution is done via rumor mongering too
- Various additional tricks to deal with removals (death certificates), etc.

Random networks

- Note that gossiping nodes pick another node in each cycle: they do not need to know all the nodes
- The actual communication pattern defines a random graph
  - by looking at these graphs, we can understand the properties of the communication better
  - we can design better gossip protocols if we understand the implications of our design decisions

The Erdős-Rényi model

- Often used to reason about gossip protocols
- Simple undirected graph $G_{N,p}$
- Parameters
  - N: number of nodes
  - p: probability of connecting any pairs of nodes
- Algorithm
  - Start with empty graph of N nodes
  - Draw all N(N-1)/2 possible edges with probability p
- Stats of degree of a fixed node i
  - $<k_i>=p(N-1)$, k_i has binomial distr, approx Poisson

Connectivity

- For gossip this is critical: can we reach all nodes using a given communication pattern?
- Let's look at connectivity as a function of p
  - AKA "graph evolution": when we keep adding edges
- Note that if p grows slower than 1/N, the graph is a disconnected collection of small (constant size) components
- If p~1/N, avg node degree <k> is constant, cycles of all order have finite probability
  - What's going on if <k> is constant?

The case when p~1/N

- 0< <k> <1
  - One cycle, otherwise trees, the larges being O(ln N) size
  - The number of clusters is N-n (ie each new edge connects two clusters)
- <k>=1
  - Critical value: largest cluster is suddenly $O(N^{2/3})$, with complex structure
- <k> >1
  - The largest cluster is of size (1-f(<k>))N nodes where f decreases exponentially
- If <k> >= ln N, completely connected (but here the avg degree grows with N)
  - Does this mean we need O(log N) neighbors to gossip to? In other words, is this a good model?

Degree distribution

- k_i the degree of fixed node
  - k_i is binomial (Bin(N-1,p))
- Degree distribution: the degree of a random node from a random graph
  - x_k: number of nodes with degree k
  - $\langle x_k \rangle = NP(k_i = k)$
  - Distribution of x_k has very low variance
  - So it is a reasonable assumption to say that a random graph $G_{N,p}$ has very close to binomial degree distribution

Diameter

- Directly related to gossip dissemination speed (ie, it is a lower bound)
- The longest shortest path
- $L = \ln N / \ln \langle k \rangle = \log_{\langle k \rangle} N$
- The intuitive reason is that these graphs are locally like trees
- The average path length (l) grows also as $\log_{\langle k \rangle} N$

Other interesting models

- $G_{r\text{-}reg}$: probability space is the set of r-regular graphs with equal probability
  - $G_{3\text{-}reg}$ is Hamiltonian
  - Note that $G_{3/(N-1),N}$ is not even connected
- $G_{r\text{-}out}$: we generate a random graph by adding 3 edges from all nodes
  - $G_{4\text{-}out}$ is Hamiltonian
  - It is believed that $G_{3\text{-}out}$ is also Hamiltonian
- Diameter is still O(log N)

Conclusions for gossiping?

- The ER model is often used to reason about gossip protocol design. This is problematic for a number of reasons
  - In the ER model nodes can get stuck without neighbors. This is the main reason for disconnectivity. In push or push-pull this is impossible
  - If we guarantee that all nodes communicate to at least 4 other nodes after receiving the update, we have a radically different model
- Message and node failure pushes the underlying network toward the ER model, but not completely

Spacial Gossip

- So far: random contacts
  - This is not optimal for underlying network traffic
  - Need to take proximity into account
- Spacial gossip: peer selection is biased according to distance of the peer: selecting node i is proportional to d^{-a} where d is the distance of i
- If the underlying topology is linear, then the expected traffic per link per cycle:

$$T(n) = \begin{cases} O(n), & a < 0; \\ O(n/\log n), & a = 1; \\ O(n^{2-a}), & 1 < a < 2; \\ O(\log n), & a = 2; \\ O(1), & a > 2 \end{cases}$$

Spatial Gossip

- a=2 is the best
  - Best tradeoff between speed and traffic
  - Probability is proportional to $1/d^2$
- Generalize to non-linear case
  - Q(d): cumulative number of sites at most at distance d
  - Probability proportional to $1/Q(d)^2$
- Smoothing out pathological topologies
  - Order all sites according to distance
  - Treat it as a linear structure

References

- Alan Demers, Dan Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard Sturgis, Dan Swinehart, and Doug Terry. Epidemic algorithms for replicated database maintenance. In Proceedings of the 6th Annual ACM Symposium on Principles of Distributed Computing (PODC'87), pages 1–12, Vancouver, British Columbia, Canada, August 1987. ACM Press.
- Eugster, P. T., Guerraoui, R., Kermarrec, A.-M., and Massoulie, L. 2004. Epidemic information dissemination in distributed systems. IEEE Computer 37, 5 (May), 60–67.
- A.-M. Kermarrec and M. van Steen, editors, Special issue of ACM SIGOPS Operating Systems Review on Gossip Protocols, 41(5), 2007.
- Boris Pittel. On spreading a rumor. SIAM Journal on Applied Mathematics, Vol. 47, No. 1 (Feb., 1987), pp. 213-223
- R Karp, C Schindelhauer, S Shenker, B Vocking. Randomized Rumor Spreading. FOCS 2000.

A Gossip Skeleton

- Originally for information dissemination in a very simple but efficient and reliable way
- Later the gossip approach has been generalized resulting in many local probabilistic and periodic protocols
- we will introduce a simple common skeleton and look at
  - information dissemination
  - topology construction
  - aggregation

A Gossip Skeleton

- the push-pull model is sown
- the active thread initiates communication (push) and receives peer state (pull)
- the passive thread mirrors this behavior

```
do once in each T time units at
a random time
    p = selectPeer()
    send state to p
    receive state_p from p
    state = update(state_p)
```

active thread

```
do forever
    receive state_p from p
    send state to p
    state = update(state_p)
```

passive thread

Rumor mongering as an instance

- state: set of active updates
- selectPeer: a random peer from the network
  - very important component, we get back to this soon
- update: add the received updates to the local set of updates
- propagation of one given update can be limited (max k times or with some probability, as we have seen, etc)

Peer Sampling

- A key method is selectPeer in all gossip protocols (influences performance and reliability)
- In earliest works all nodes had a global view to select a random peer from
  - scalability and dynamism problems
- Scalable solutions are available to deal with this
  - random walks on fixed overlay networks
  - dynamic random networks

Random walks on networks

- if we are given any fixed network, we can sample the nodes with any arbitrary distribution with the Metropolis algorithm:

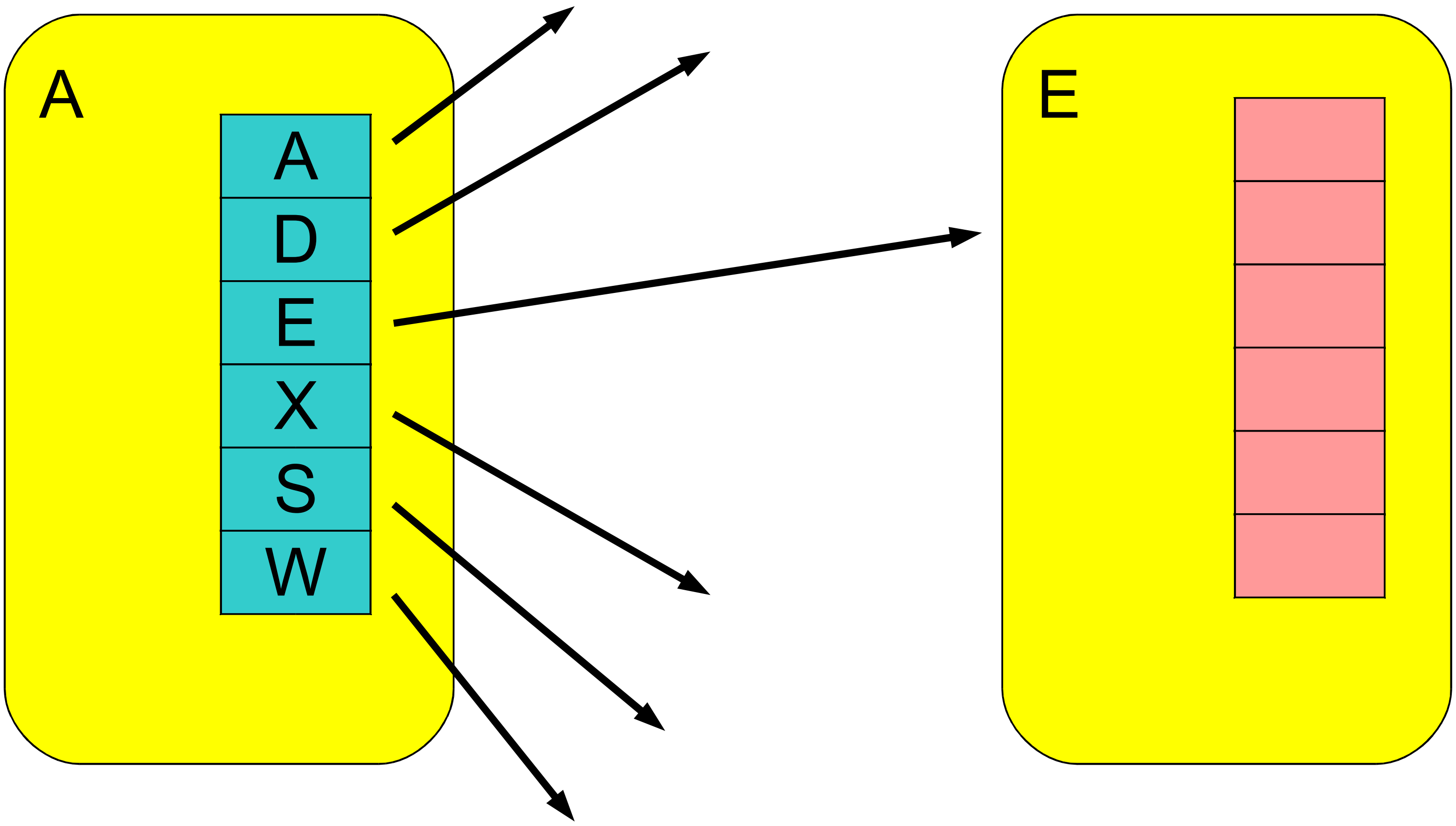
$$P_{i,j} = \begin{cases} \frac{1}{2} \cdot \frac{1}{d_i} & \text{if } \frac{\pi(i)}{d_i} \leq \frac{\pi(j)}{d_j}; \\ \frac{1}{2} \cdot \frac{1}{d_j} \cdot \frac{\pi(j)}{\pi(i)} & \text{if } \frac{\pi(i)}{d_i} > \frac{\pi(j)}{d_j}. \end{cases}$$

- This Markov chain has stationary distribution π where d_i is the degree of node i (undirected graph)

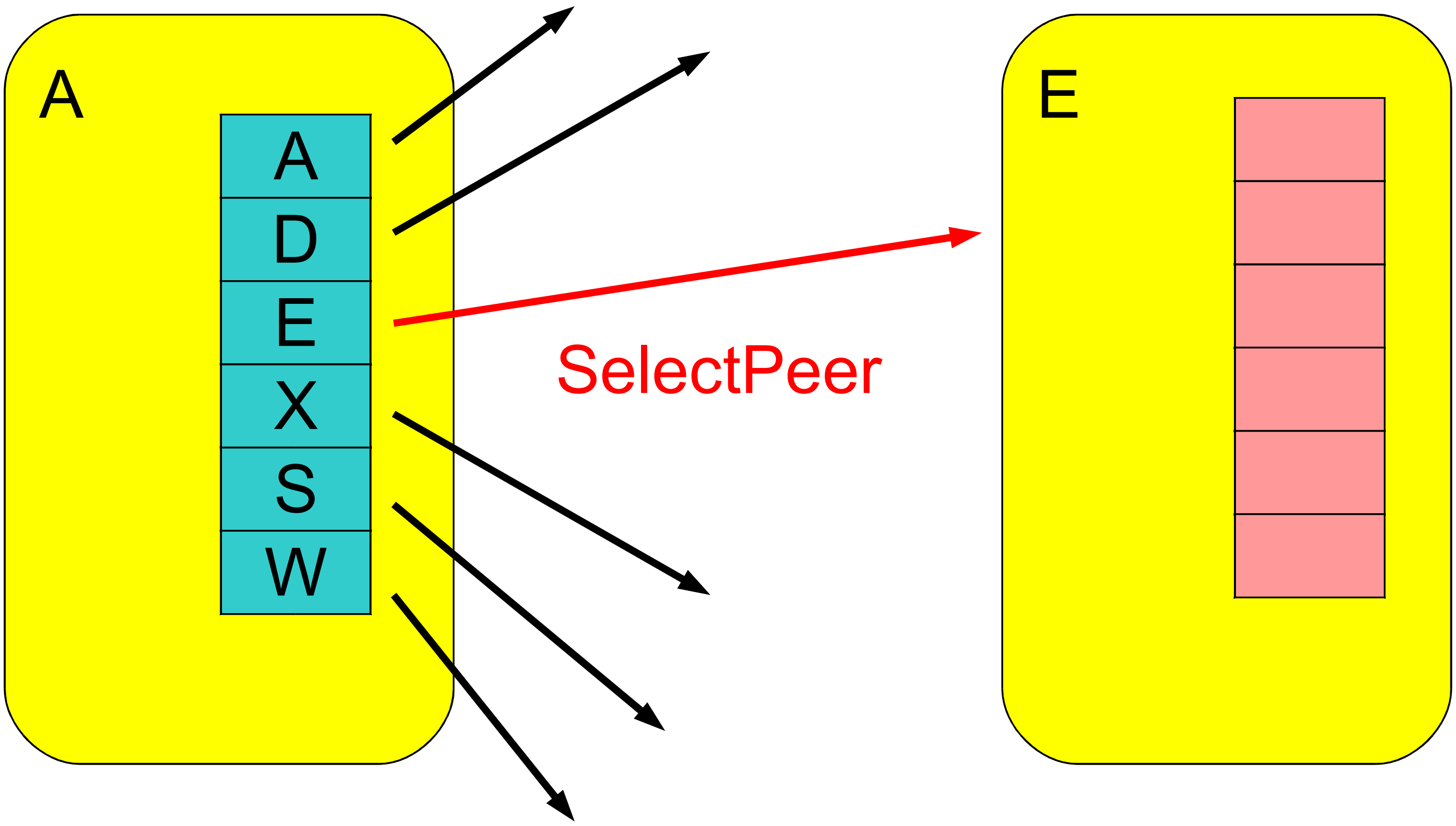
Gossip based peer sampling

- basic idea: random peer samples are provided by a gossip algorithm: the peer sampling service
- The peer sampling service uses itself as peer sampling service (bootstrapping)
  - no need for fixed (external) network
- state: a set of random overlay links to peers
- selectPeer: select a peer from the known set of random peers
- update: for example, keep a random subset of the union of the received and the old link set

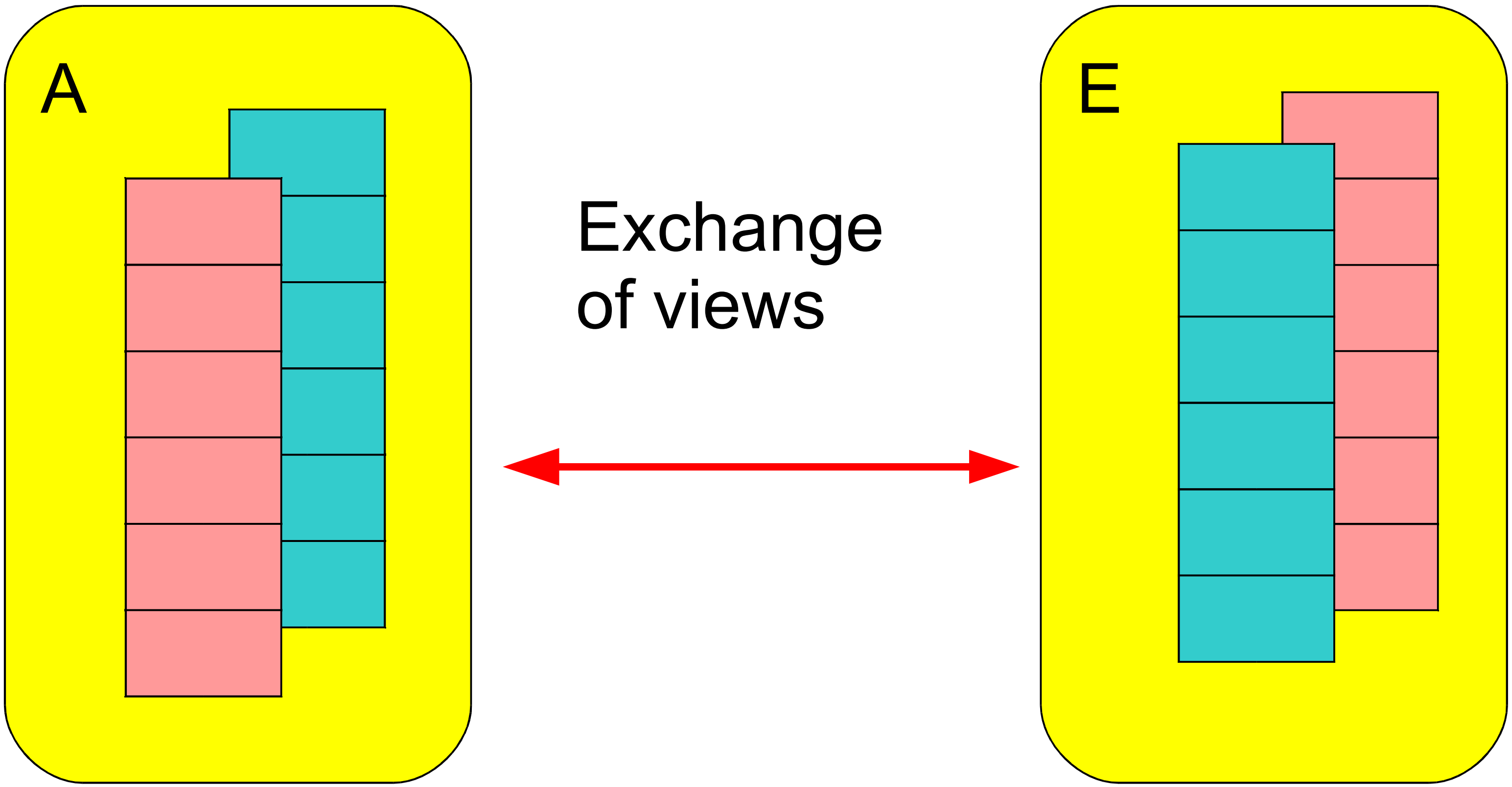
Gossip protocols for topology management



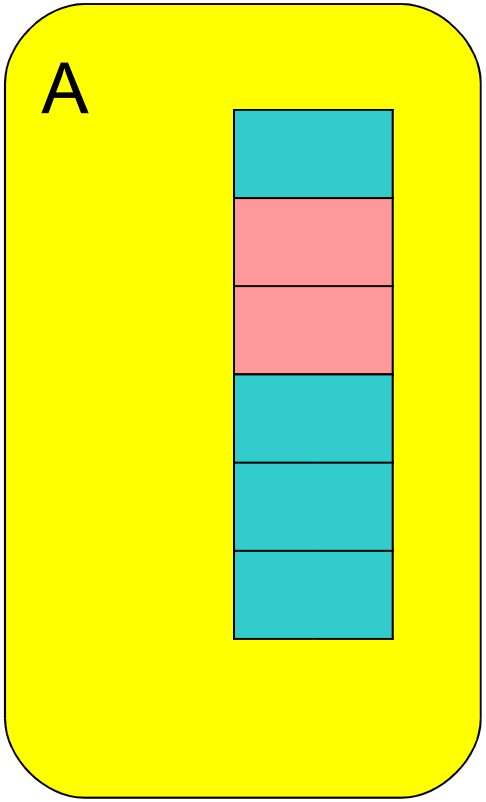
Gossip protocols for topology management



Gossip protocols for topology management

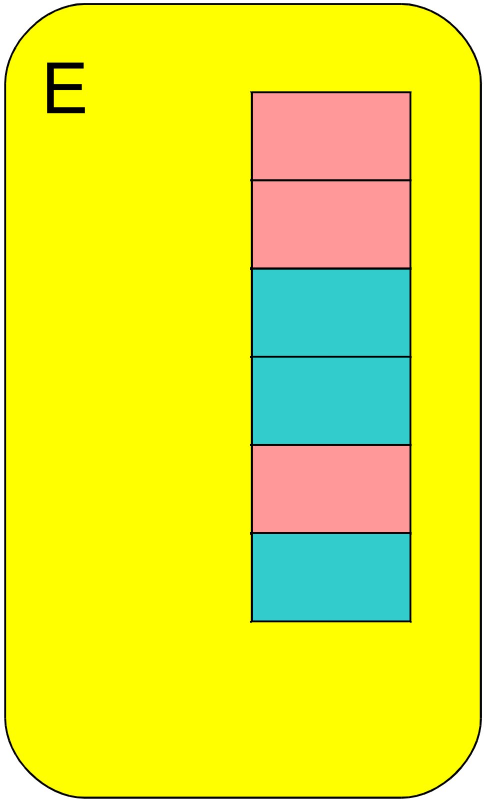


Gossip protocols for topology management



Both sides apply update

thereby redefining topology



Gossip based peer sampling

- in reality a huge number of variations exist
  - timestamps on the overlay links can be taken into account: we can select peers with newer links, or in update we can prefer links that are newer
- these variations represent important differences w.r.t. fault tolerance and the quality of samples
  - the links at all nodes define a random-like overlay that can have different properties (degree distribution, clustering, diameter, etc)
  - turns out actually not really random, but still good for gossip

References

- Allavena, A., Demers, A., and Hopcroft, J. E. 2005. Correctness of a gossip based membership protocol. In Proceedings of the 24th annual ACM symposium on principles of distributed computing (PODC 05). ACM Press, Las Vegas, Nevada, USA.
- Márk Jelasity, Rachid Guerraoui, Anne-Marie Kermarrec, and Maarten van Steen. The peer sampling service: Experimental evaluation of unstructured gossip-based implementations. In Hans-Arno Jacobsen, editor, Middleware 2004, volume 3231 of Lecture Notes in Computer Science, pages 79–98. Springer-Verlag, 2004. (journal version: ACM TOCS 2007 aug)
- Zhong, M., Shen, K., and Seiferas, J. 2005. Non-uniform random membership management in peer-to-peer networks. In Proc. of the IEEE INFOCOM. Miami, FL.

Gossip based topology management

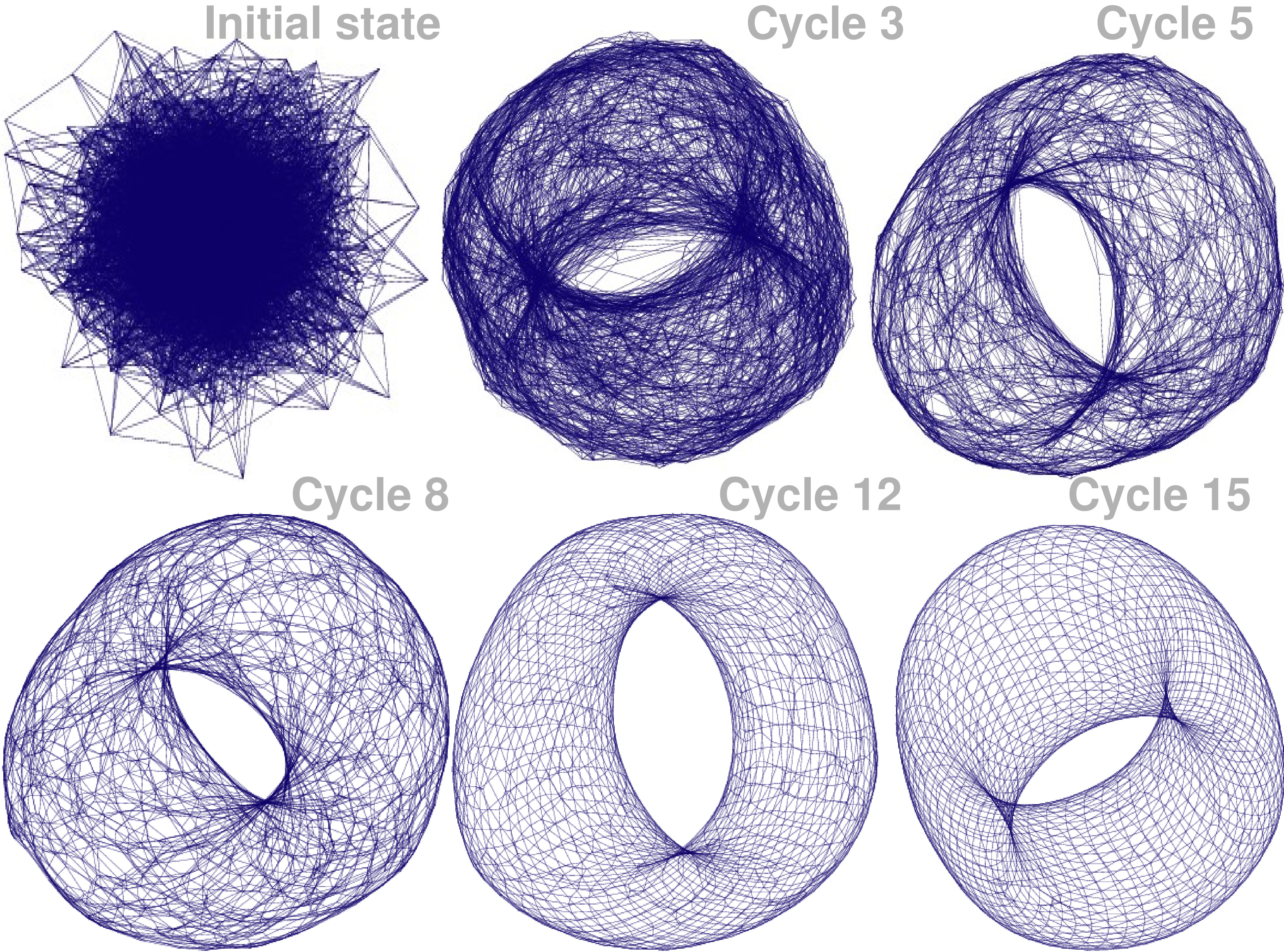
- We saw we can build random networks. Can we build any network with gossip?
- Yes, many examples
  - proximity networks
  - DHT-s (Bamboo DHT: maintains Pastry structure with gossip inspired protocols)
  - semantic proximity networks
  - etc

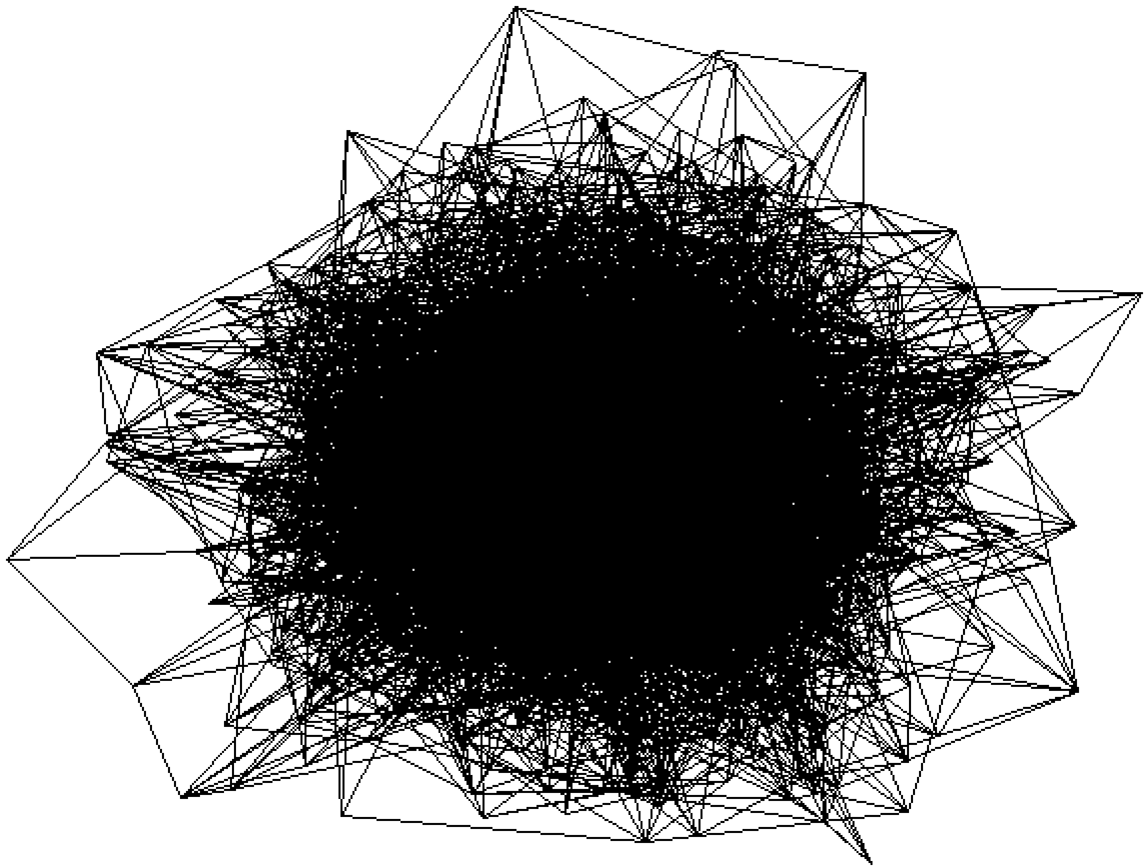
T-Man

- T-MAN is a protocol that captures many of these in a common framework, with the help of the ranking method:
  - ranking is able to order any set of nodes according to their desirability to be a neighbor of some given node
  - for example, based on hop count in a target structure (ring, tree, etc)
  - or based on more complicated criteria not expressible by any distance measure

Gossip based topology management

- basic idea: random peer samples are provided by a gossip algorithm: the peer sampling service
- state: a set of overlay links to peers
- selectPeer: select the peer from the known set of peers that ranks highest according to the ranking method
- update: keep those links that point to nodes that rank highest


Initial state
Cycle 3
Cycle 5
Cycle 8
Cycle 12
Cycle 15



References

- Márk Jelasity and Ozalp Babaoglu. T-Man: Gossip-based overlay topology management. In Sven A. Brueckner, Giovanna Di Marzo Serugendo, David Hales, and Franco Zambonelli, editors, Engineering Self-Organising Systems: Third International Workshop (ESOA 2005), Revised Selected Papers, volume 3910 of Lecture Notes in Computer Science, pages 1–15. Springer-Verlag, 2006.
- Sean Rhea, Dennis Geels, Timothy Roscoe, and John Kubiatowicz. Handling Churn in a DHT. Proceedings of the USENIX Annual Technical Conference, June 2004.
- Laurent Massoulie, Anne-Marie Kermarrec, and Ayalvadi J. Ganesh. Network awareness and failure resilience in self-organising overlay networks. In Proceedings of the 22nd Symposium on Reliable Distributed Systems (SRDS 2003), pages 47–55, Florence, Italy, 2003.
- Spyros Voulgaris and Maarten van Steen. Epidemic-style management of semantic overlays for content-based searching. In Jose C. Cunha and Pedro D.Medeiros, editors, Proceedings of Euro-Par, number 3648 in Lecture Notes in Computer Science, pages 1143–1152. Springer, 2005.

Aggregation

- Calculate a global function over distributed data
  - eg average, but more complex examples include variance, network size, model fitting, etc
- usual structured/unstructured approaches exist
  - structured: create an overlay (eg a tree) and use that to calculate the function hierarchically
  - unstructured: design a stochastic iteration algorithm that converges to what you want (gossip)
- we look at gossip here

Implementation of aggregation

- state: current approximation of the average
  - initially the local value held by the node
- selectPeer: a random peer (based on peer sampling service)
- updateState(s_1,s_2)
  - $(s_1+s_2)/2$: result in averaging
  - $(s_1 s_2)^{1/2}$: results in geometric mean
  - $\max(s_1,s_2)$: results in maximum, etc

Illustration of averaging

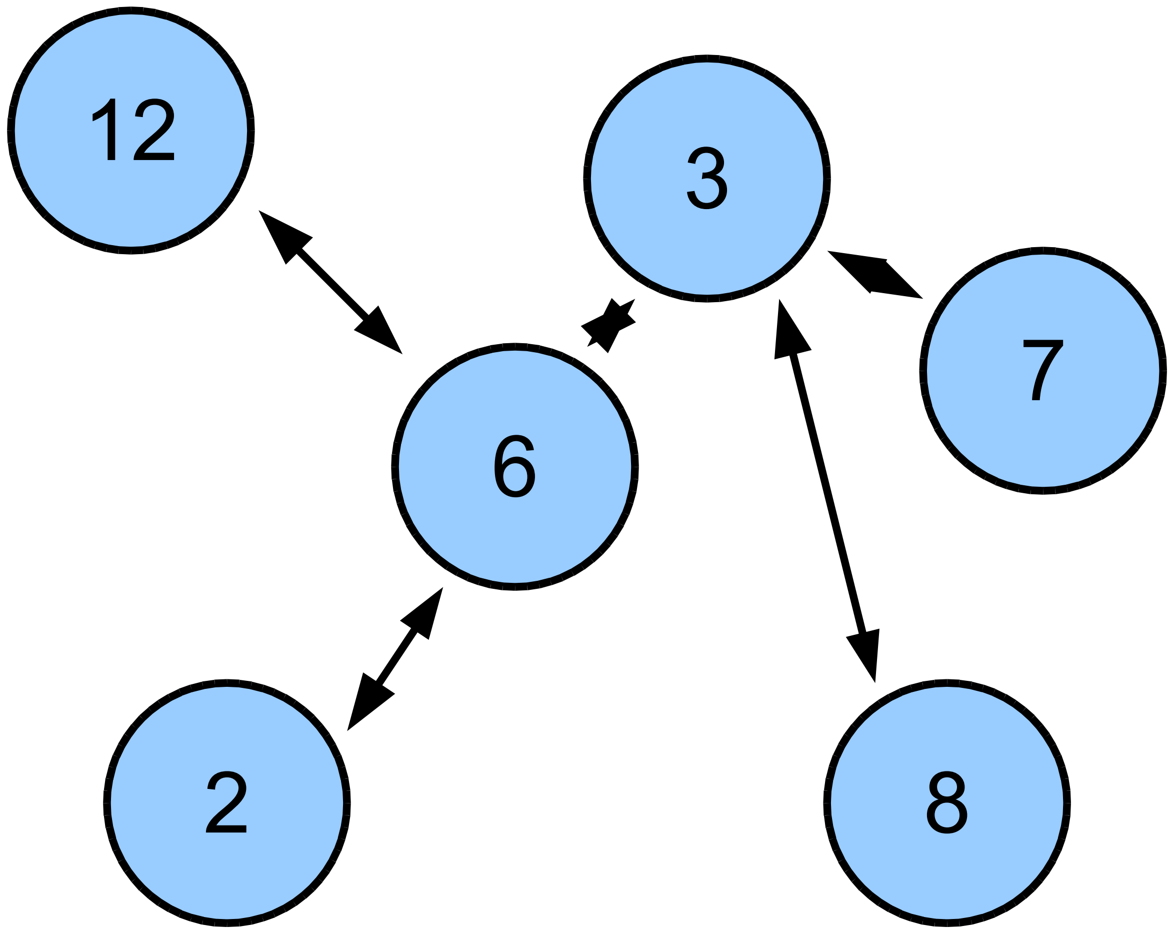
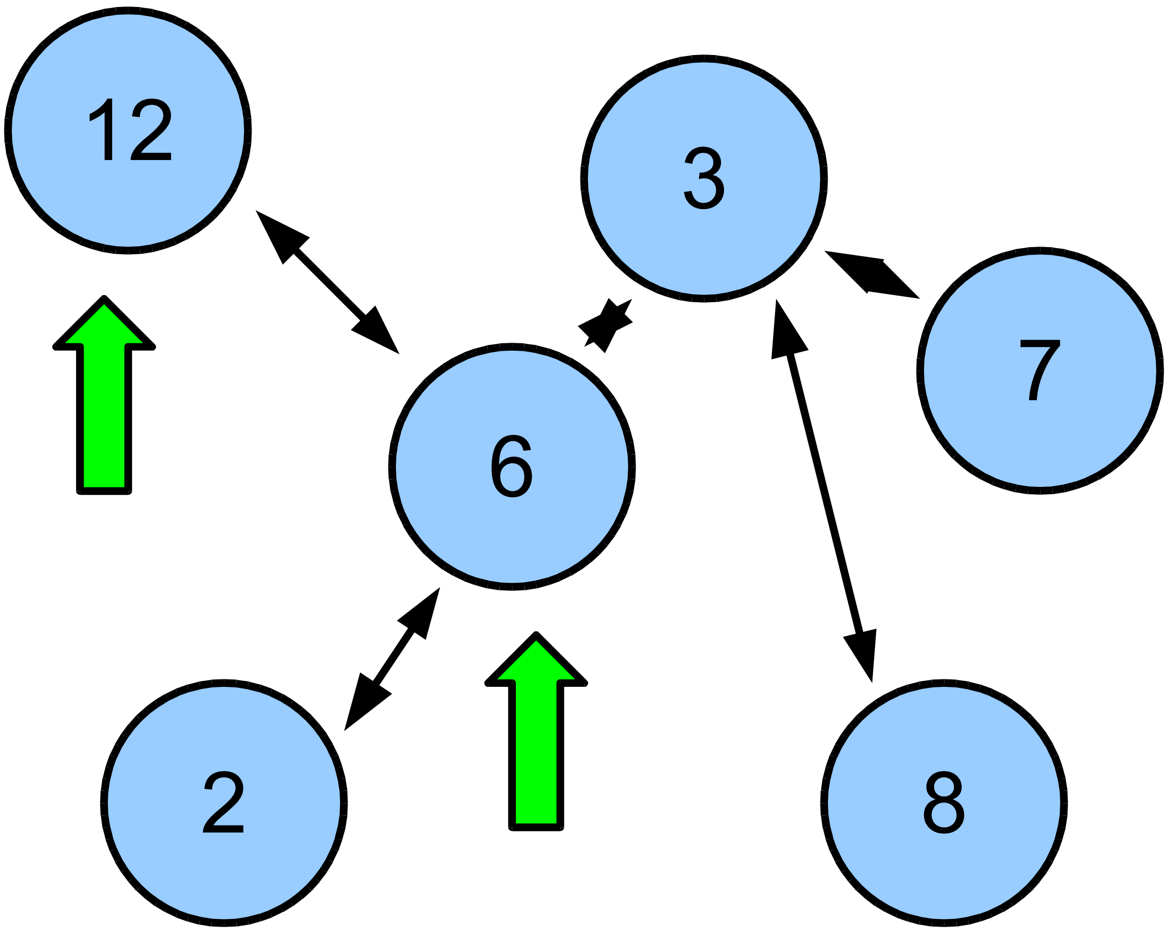
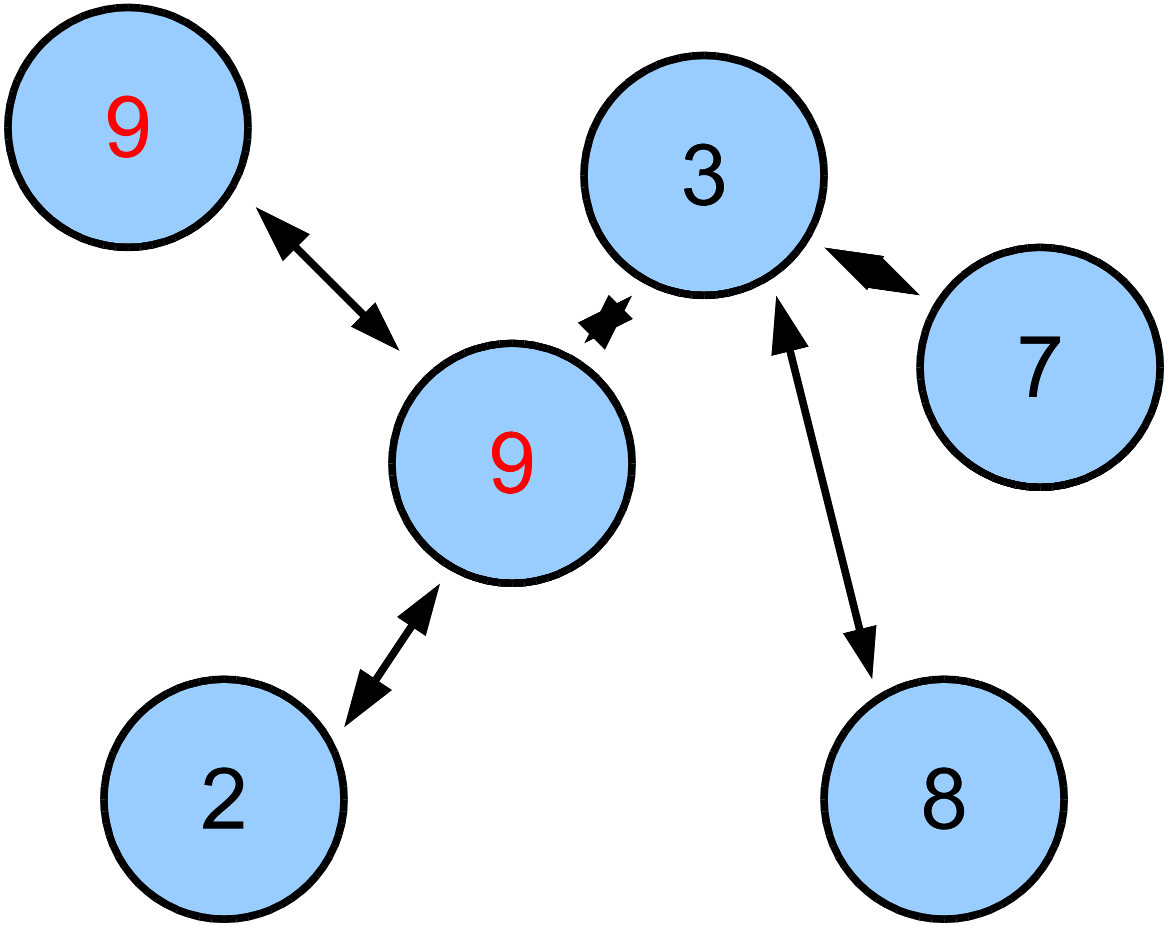


Illustration of averaging



(12+6)/2=9

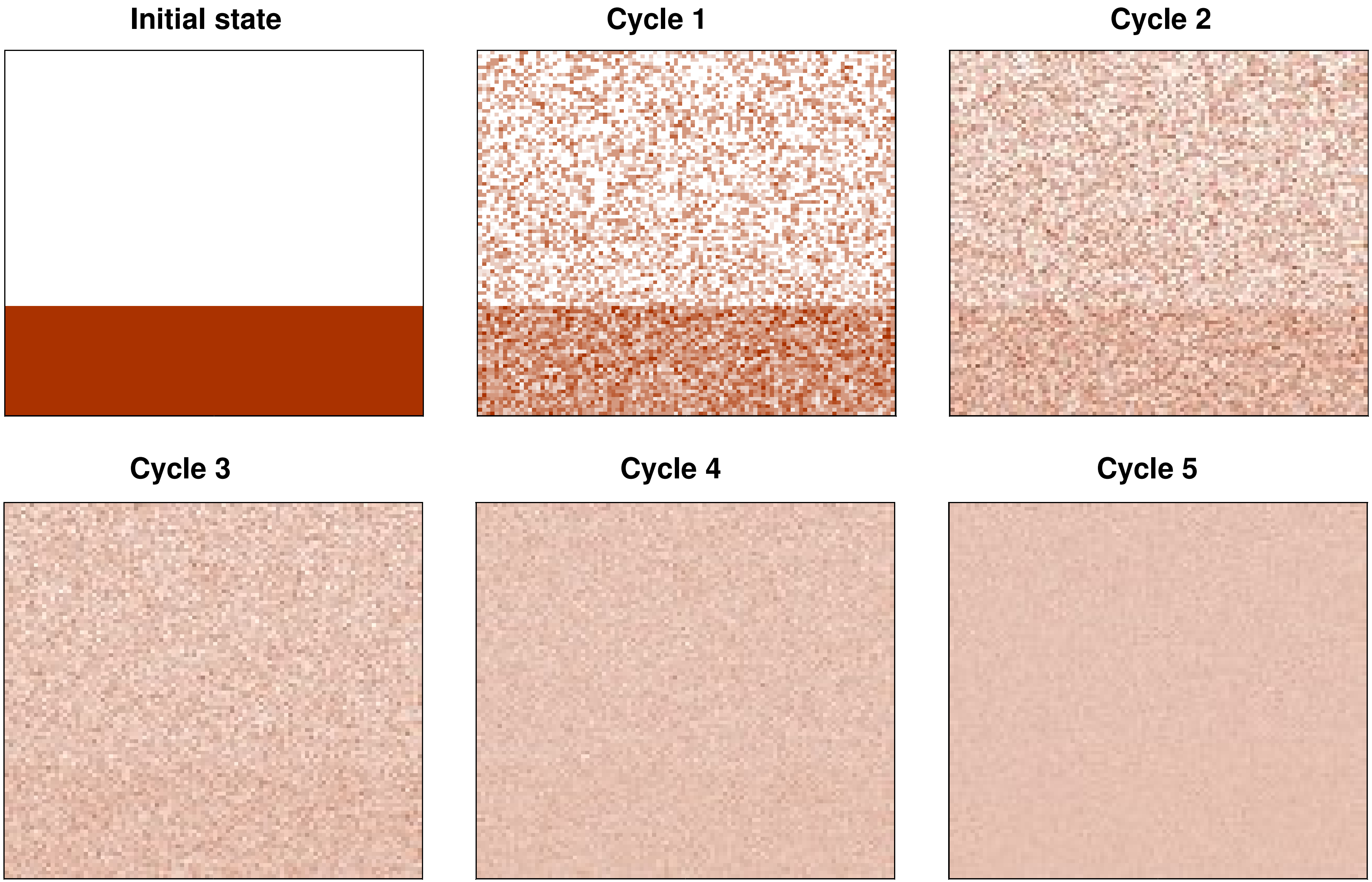
Illustration of averaging



Improvements

- Tolerates asymmetric message loss (only push or pull) badly
- Tolerates overlaps in pairwise exchanges badly
- [Kempe et al 2003] propose a slightly different version
  - all nodes maintain s (sum estimate) and w (weight)
  - estimate is s/w
  - only push: send (s/2,w/2), and keep s=s/2, w=w/2
- several other variations exist

Illustration of averaging



References

- Kempe, D., Dobra, A., and Gehrke, J. 2003. Gossip-based computation of aggregate information. In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS’03). IEEE Computer Society, 482–491.
- Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. Gossip-based aggregation in large dynamic networks. ACM Transactions on Computer Systems, 23(3):219–252, August 2005.
- Robbert van Renesse, Kenneth P. Birman, and Werner Vogels. Astrolabe: A robust and scalable technology for distributed system monitoring, management, and data mining. ACM Transactions on Computer Systems, 21(2):164–206, May 2003.

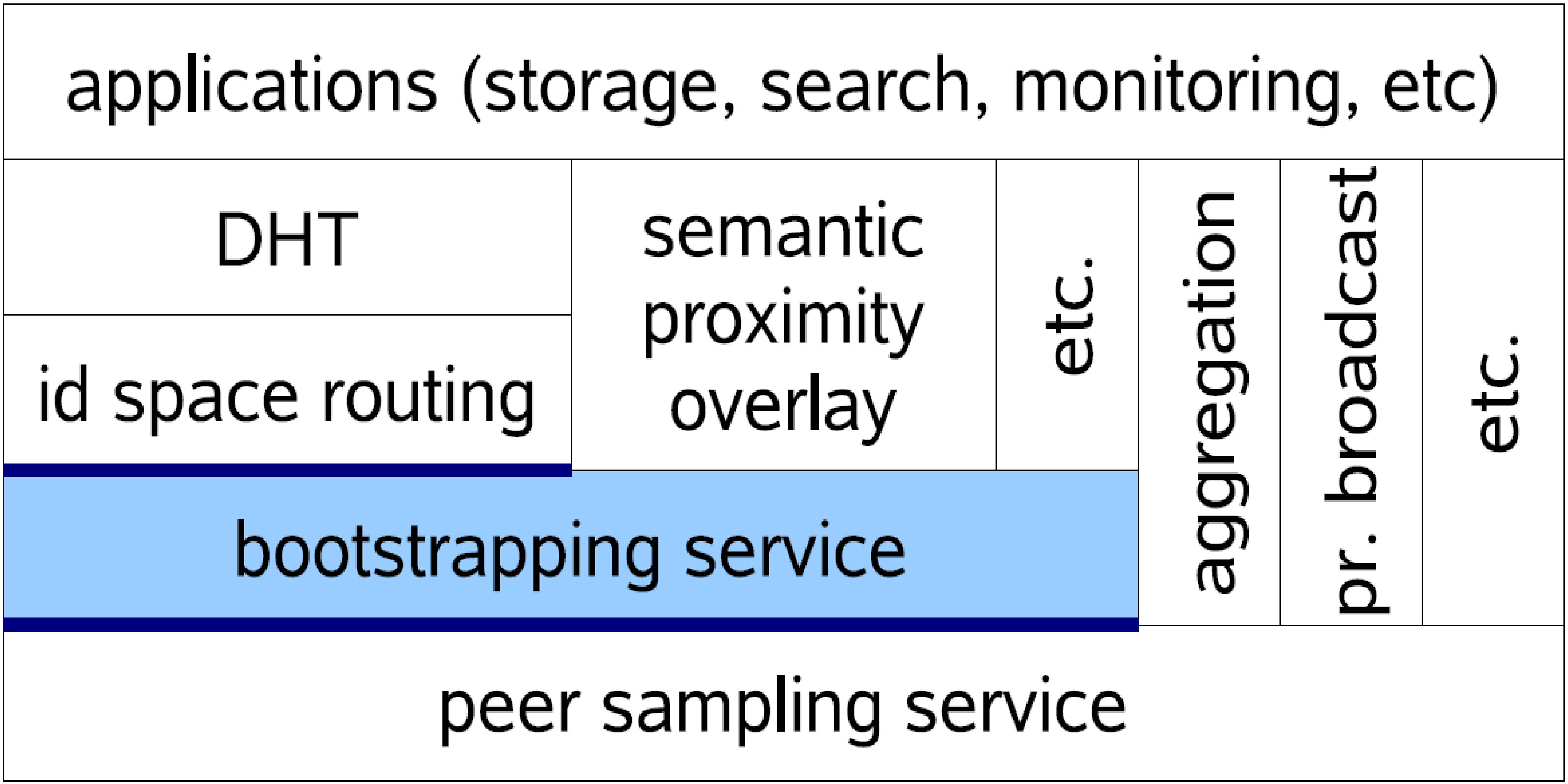
Some other examples

- firefly-inspired synchronization
- partitioning (slicing) and sorting in P2P networks
- asynchronous implementation of matrix iterations
  - ranking (PageRank)
  - reputation systems
- emergent cooperation

Modular design

- We have seen that all gossip protocols use the peer sampling service that is itself a gossip protocol
- Can be generalized: gossip protocols can be stacked or arbitrarily combined
  - actual local communication is the same (all protocols can often piggyback the same message)
  - conceptual structure is modular

Example modular architecture



Outlook

- Gossip is similar to many other fields of research that also have some of the following features:
  - periodic, local, probabilistic, symmetric
- examples include
  - swarm systems, cellular automata, parallel asynchronous numeric iterations, self-stabilizing protocols, etc

A slide on viruses and worms

- We focused on “good” epidemics but malicious applications are known
  - viruses and worms replicate themselves via similar algorithms using some underlying network such as email contacts or the Internet itself
- The dynamics is described by SIS model
- Underlying networks are typically scale free (power law degree distribution)
  - can be proven: no threshold: it is nearly impossible to completely eliminate a “disease”

Some open problems

- gossip in mobile contact networks and its potential applications (also malware...)
- security
  - gossip is robust to benign failure but very sensitive to malicious attacks
  - current "secure" gossip protocols sacrifice simplicity and light-weight
- interdisciplinary connections: toward a deeper understanding of self-organization and gossip protocols as a special case