Discovering Information in Complex Networks

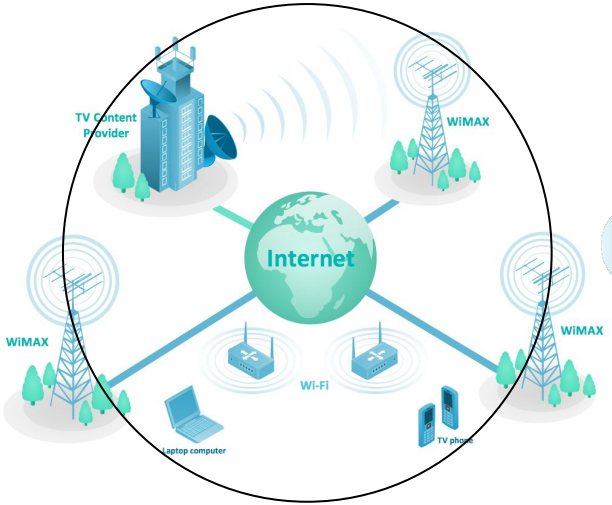
Liang Wang, Suzan Bayhan, Jörg Ott, Jussi Kangasharju, Arjuna Sathiaseelan, Jon Crowcroft

University of Cambridge, UK Aalto University, Finland University of Helsinki, Finland

liang.wang@cl.cam.ac.uk

What Is the Jigsaw in My Research?

Information



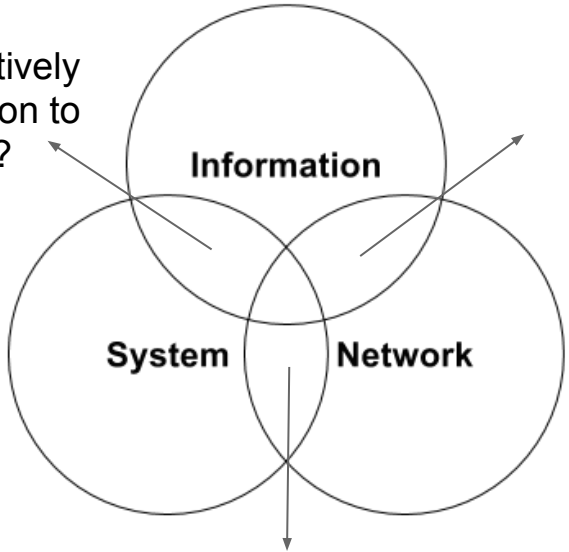
Network

System

UNIVERSITY OF CAMBRIDGE

What Happens When One Meets Another?

Information System
How to build a system to effectively manage and analyse information to facilitate knowledge discovery?

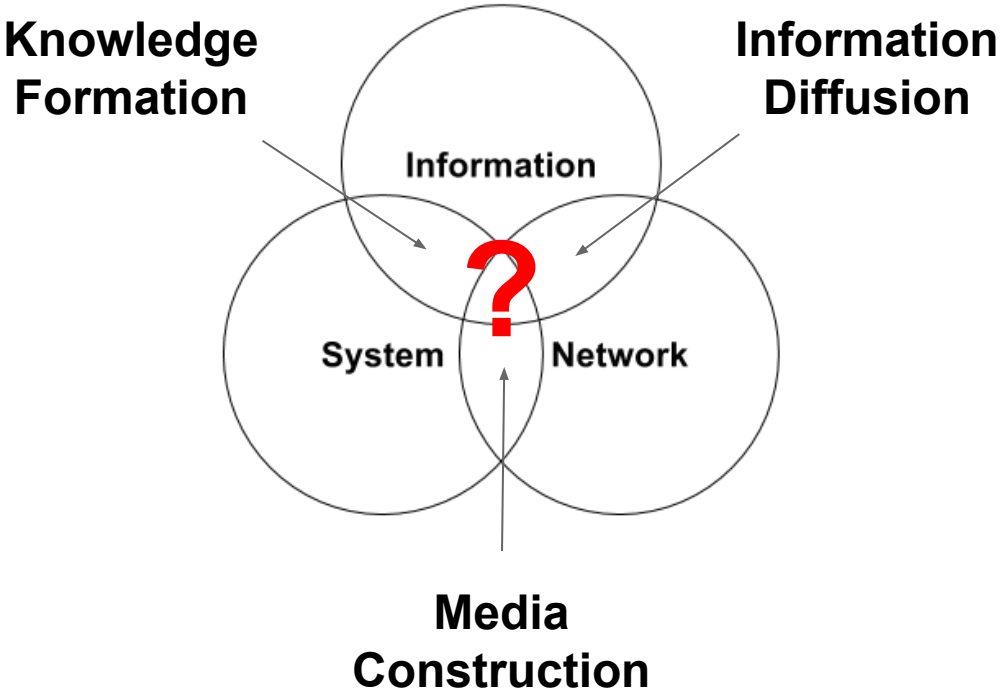


Information-Centric Networking
How to build a communication protocol to efficiently distribute information in a network?

Network Operating System
How to build an operating system to adaptively manage networked resources under various dynamics?

UNIVERSITY OF CAMBRIDGE

What Sits in the Center?



UNIVERSITY OF CAMBRIDGE

What A Cool System!

?

So, we have a system …

which can make sense of data → information → knowledge,

which can adapt itself under various network dynamics,

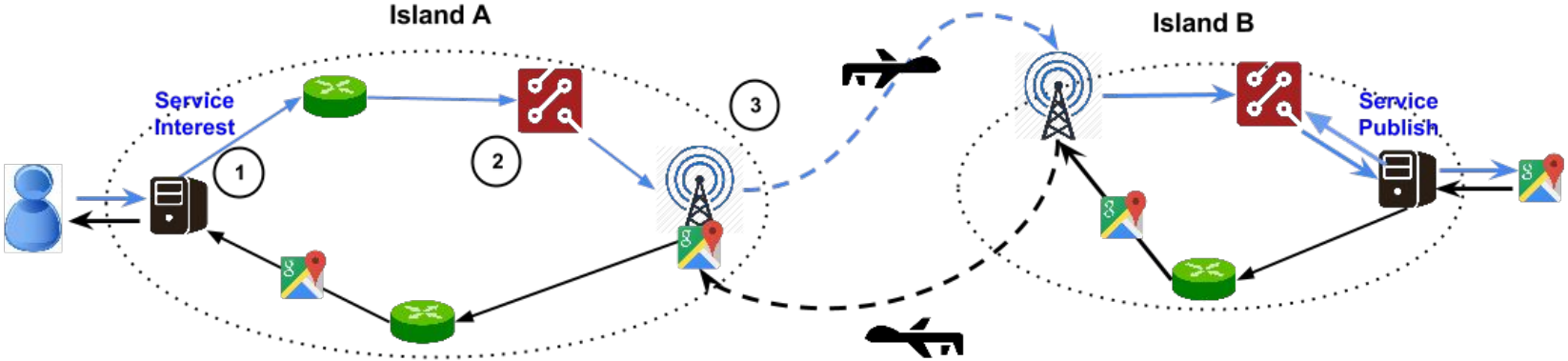
which can deliver knowledge safely and efficiently to

wherever and whenever they are needed …

What A Great Vision!

!

Diffusing knowledge, as far as possible ...

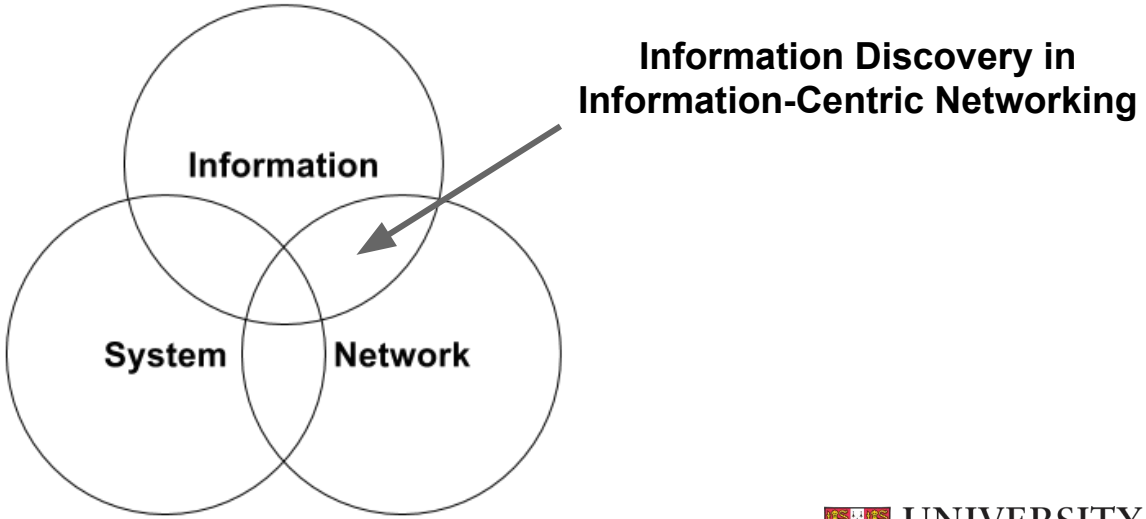


UNIVERSITY OF CAMBRIDGE

What Is Next?

OK, warm-up ends.

Today, we only focus on ...



What Do We Actually Want to Study?

- Benefits of (scoped) flooding in the network
  - Content discovery, routes propagation, etc.
  - Low state maintenance, low protocol complexity, etc.
  - A scalable solution or not?

- Technically we want to know
  - How to set the flooding scope optimally?
  - How a network topology impacts the scope?
  - How content availability impacts the scope?

In short, we want to flood on the right content at right place with right scope.

Is This Really An Important Problem?

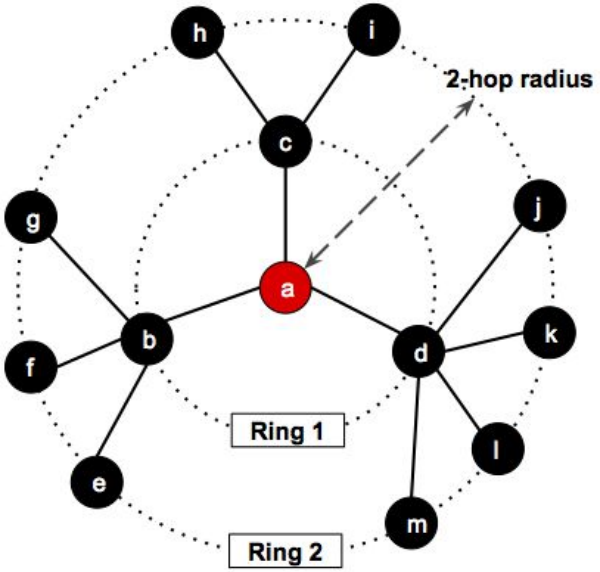
- Flooding is widely used but it lacks of theoretical backup.
- Understanding scope-flooding has further implications on other topics such as opportunistic network, P2P, and etc.
- Lack of a network model to study the neighbourhood.
- Lack of a cost/gain model to study flooding related problems. Most importantly, the model should be extendable.

What Do We Need to Start With?

- Three components are needed:
  - The content (can be anything), only its value matters.
  - The representation of gain/cost as a function of # of nodes and content (value).
  - The network model based on which, we can tell how the # of nodes increases as a function of # of hops (scope).

How Are These Components Connected?

- A node-centric ring-based model



node who initiates the flooding

node who relays the flooding

Utility value is decomposed into multiple layers accordingly. Flooding stops before the ring where utility drops below zero.

U_1 = gain(b ... d) - cost(b ... d)

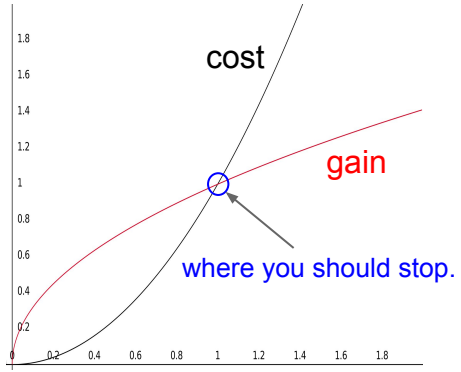
U_2 = gain(e ... m) - cost(e ... m)

UNIVERSITY OF CAMBRIDGE

How Shall We Model Gain and Cost?

- Both gain and cost are functions of # of nodes.
- Important presumption:

  After certain point, cost grows faster than gain.
- Does this presumption make sense?
  - If gain is always lower, you will never flood. Just stay still.
  - If gain always grows faster, you will never stop flooding.



How Is the Network Model Constructed?

- We use G = (V, p) instead of G = (V, E) as basis. Why?
- How fast the neighbourhood grows while the hop increases?
- Model functionality: given a scope r, the network model calculates how many nodes can we reach.
- Remember, nodes can fail, and messages can get lost.

What Can the Network Model Do?

- If we define the average network growth rate (beta) as the average ratio between # of ring r+1 nodes and # of ring r nodes,
- beta = (# of 2-hop neighbours / # of 1-hop neighbours).
- A node can estimate its neighbourhood with 2-hop knowledge.
- We considered two network generative models: Random and Scale-free networks. Both have closed-form expressions.
- What is the caveat?

How Accurate Can This Model Predict?

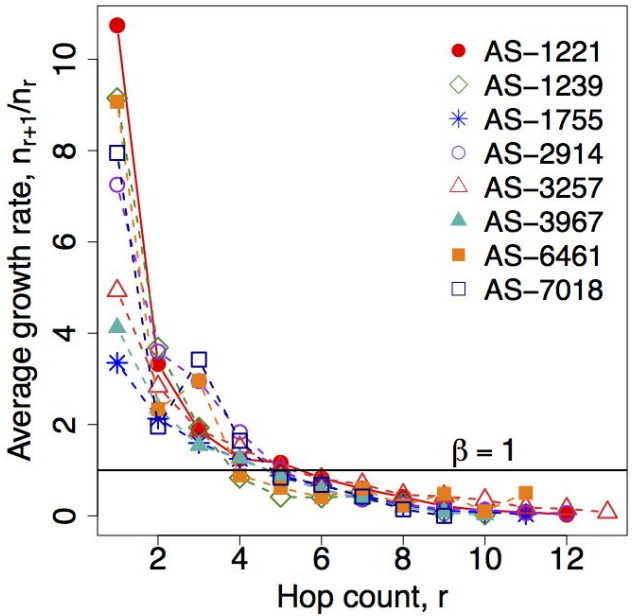
Table 1: Overestimation of the model at each hop for various network graphs. V: Number of nodes and E: Number of nodes in the generated instance of the graph, l: average path length. Shaded cells represent the cases where the error is below 0.20.

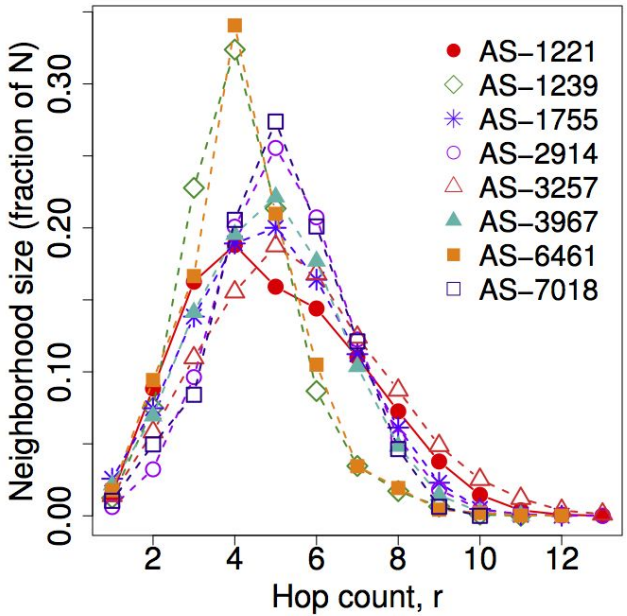
Id	Topology	V	E	$\langle k \rangle$	l	Clustering	Overestimation of the model				
							$r=2$	$r=3$	$r=4$	$r=5$	$r=6$
1	Random	339	338	1.994	23.07	0	0.327	1.046	2.359	4.692	9.092
2	Random	8030	9761	2.431	12.03	0	0.152	0.371	0.642	0.972	1.399
3	Random	9426	15068	3.197	8.30	0.00040	0.060	0.130	0.212	0.332	0.565
4	Random	9811	20073	4.091	6.75	0.00049	0.023	0.053	0.106	0.259	0.873
5	Random	9928	25060	5.048	5.88	0.00048	0.004	0.017	0.079	0.419	2.79
6	Random	9989	35020	7.011	4.95	0.00066	0.003	0.030	0.229	2.139	54.124
7	Scale-free, α =3.24	7141	9648	2.70	7.88	0.00057	0.093	0.271	0.529	1.069	2.599
8	Scale-free, α =3.35	5869	7347	2.50	8.66	0.00076	-0.115	-0.174	-0.194	-0.16	0.013
9	Scale-free, α =3.50	5960	7357	2.47	8.99	0.00013	-0.356	-0.555	-0.68	-0.757	-0.794

Pretty accurately for big networks for 3 - 4 hops.

The larger the network is, the more accurate model can predict, the reason is due to the *small network diameter.*

How Accurate Can This Model Predict?





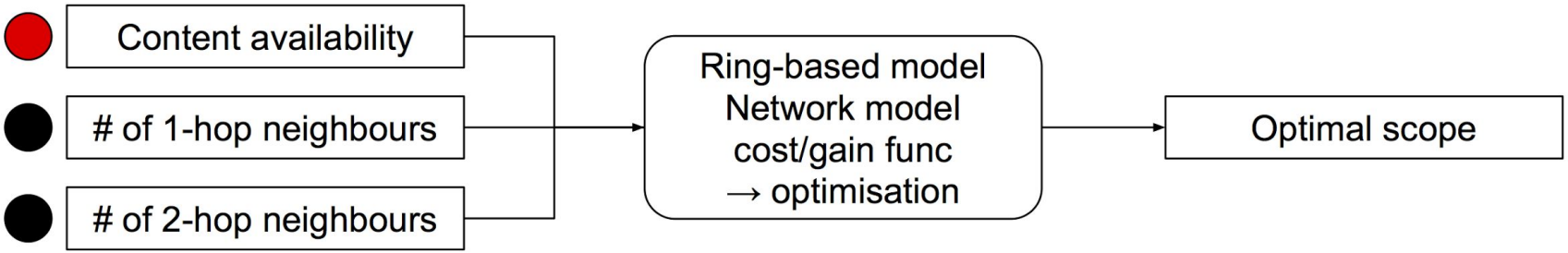
Fast growth till 4-5 hops! Then drops due to limited network diameter.

UNIVERSITY OF CAMBRIDGE

What Is the Missing Piece in Our Model?

- Do not forget the purpose of a flooding - content discovery.
- We consider two cases of a given content set.
  - The availability is given as a priori knowledge.
  - The availability is unknown, so we apply Bayesian inference to estimate.
- The rationality behind: the easier to find a content among nearby nodes, the higher its availability is.

How to Calculate the Optimal Scope?

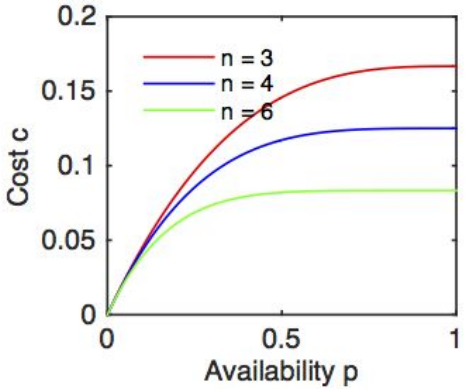


A good flooding strategy requires:

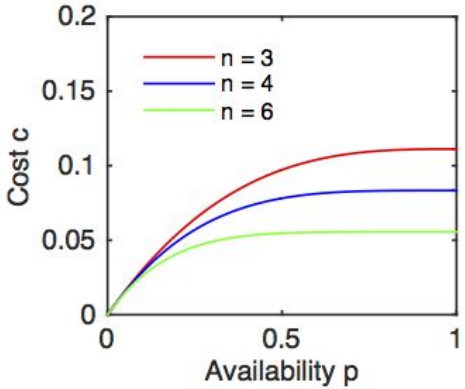
- A node is aware of its neighbourhood with an accurate topological inference.
- A node is aware of content availability with an accurate statistical inference on user request streams.

How Does the Model Behave?

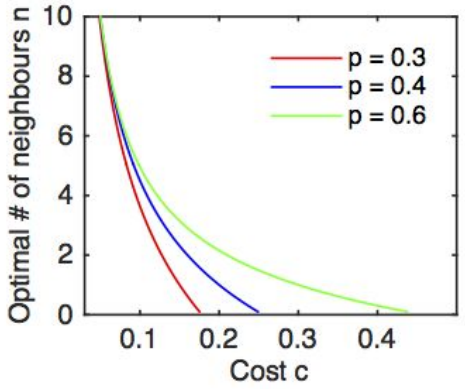
- Does the model generate meaningful behaviours?



(a) c vs. p, $\gamma = 1$.



(b) c vs. p, $\gamma = 0.5$.



(c) n vs. c, $\gamma = 1$.

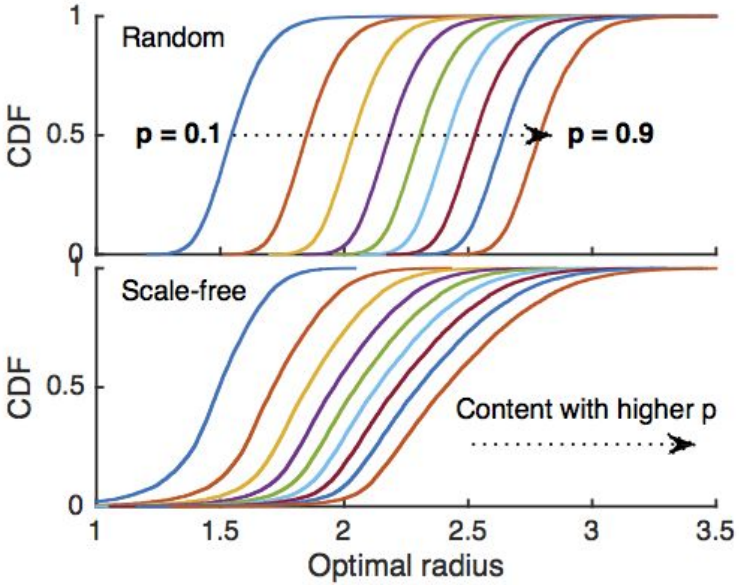
UNIVERSITY OF CAMBRIDGE

What Flooding Strategies Are Studied?

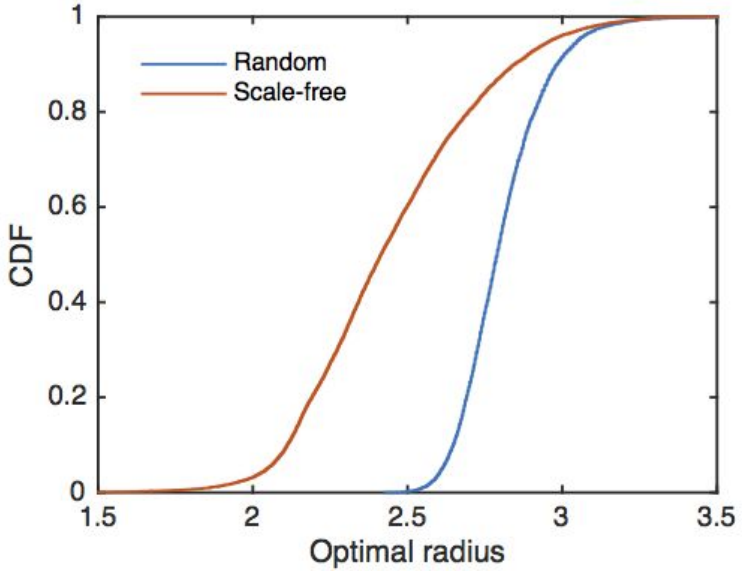
- Static Flooding (r)
  - Same optimal scope for all nodes.
  - Scope is optimised over the whole network using average # of 1-hop and 2-hop neighbours of the network.

- Dynamic Flooding (r_i for node i)
  - Scope calculated for each node: a node utilises its local (2-hop) topological information to optimise.
  - With content availability, only flood on popular content.
  - Without content availability, always flood 1-hop neighbours by default.

Do Graph Generative Models Matter?



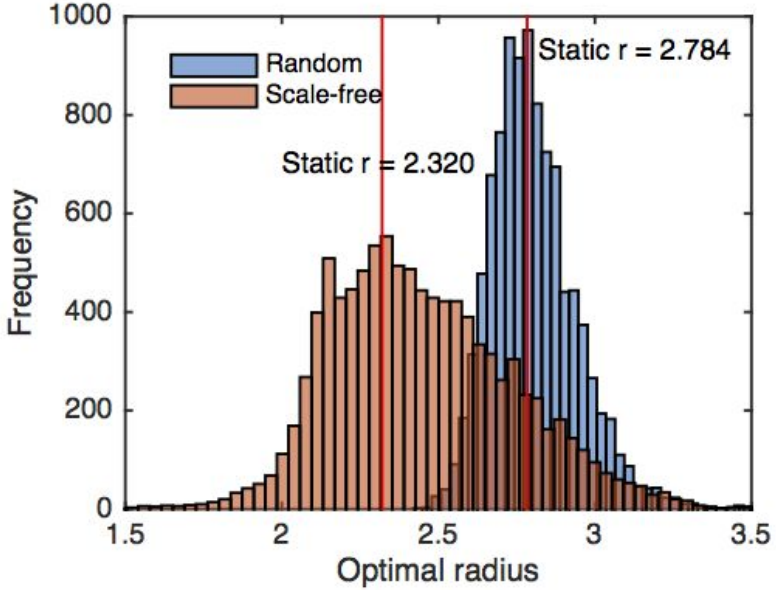
(a) CDF of radius with different p.



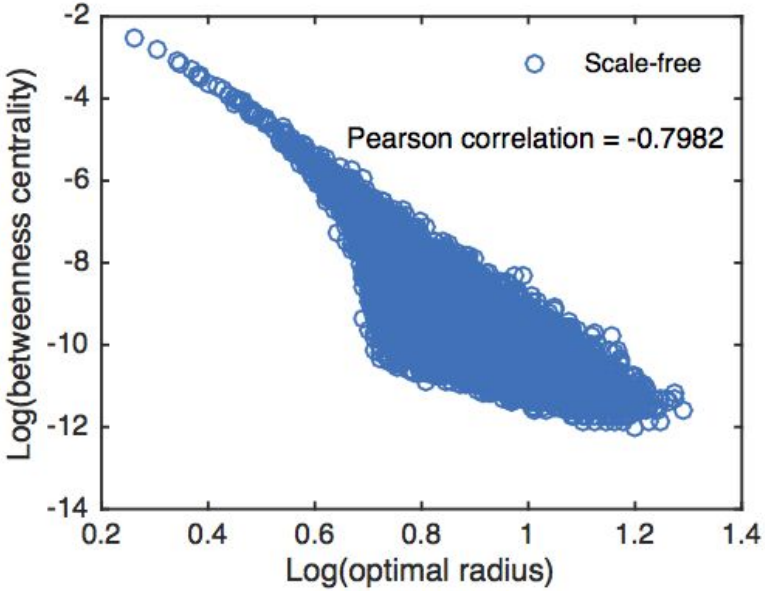
(b) CDF of optimal radius.

p: Content availability

Do Graph Generative Models Matter?



(c) Histogram of optimal radius.



(d) Radius vs. betw. centrality.

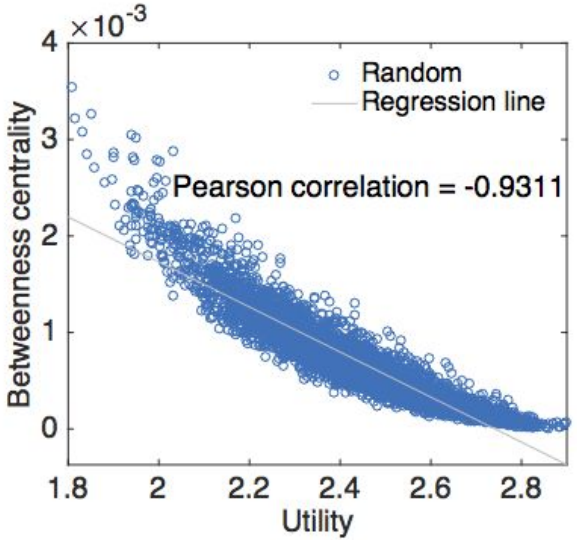
Scale free: more heterogeneity, more divergence from network wide optimal scope.

How Utilities Are Distributed in A Network?

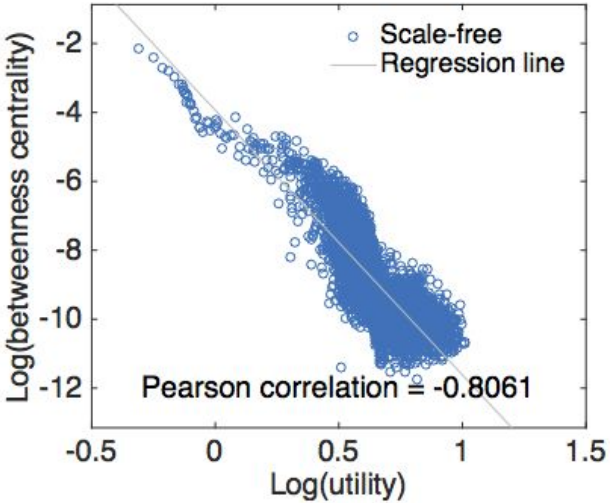
Strong negative correlation between the utility and betw. centrality.

In the dense area, a node has a high betw. centrality, it may include more neighbours than necessary (the optimum) even just for 1-hop neighbours.

The growth rate in the sparser area is lower, so nodes have a better control over the nbhd size by fine-tuning their scope leading to smaller cost and better utility.



(a) Random network

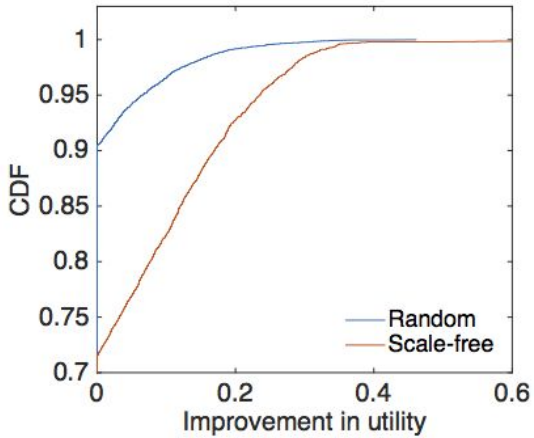


(b) Scale-free network

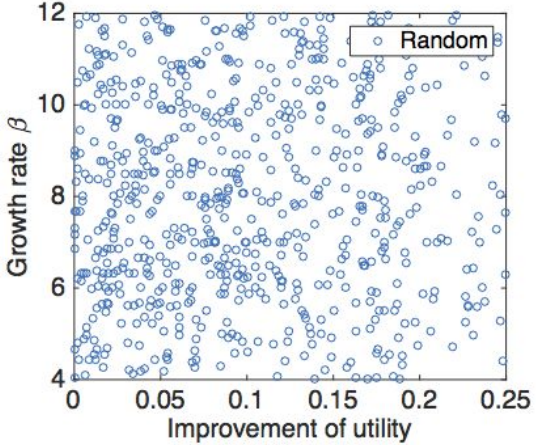
UNIVERSITY OF CAMBRIDGE

Is Dynamic Flooding Always Effective?

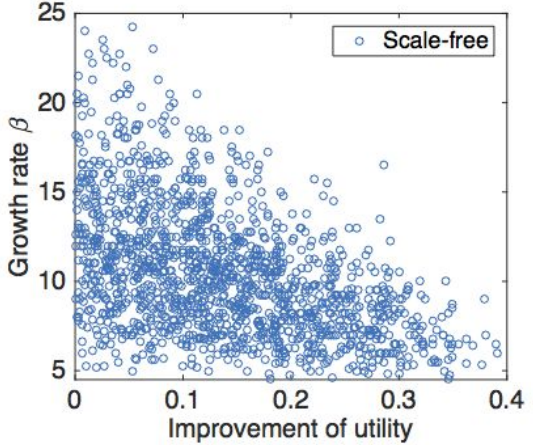
Improvement = (Utility of dynamic flooding - utility of static flooding) / utility of static flooding



(c) Dynamic against static



(d) Random network

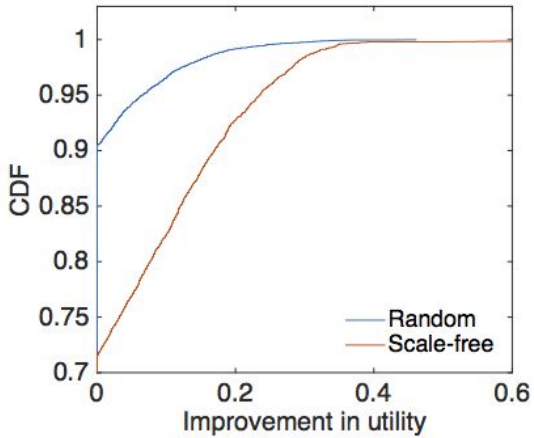


(e) Scale-free network

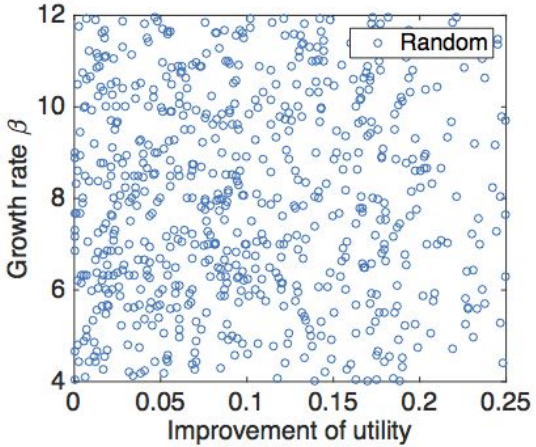
Dynamic flooding is less effective on random networks, only 10% of the nodes actually improve their performance and over half have less than 10% improvement. In scale-free network, 30% of the nodes are improved, among which over 60% have larger than 10% improvement.

Is Dynamic Flooding Always Effective?

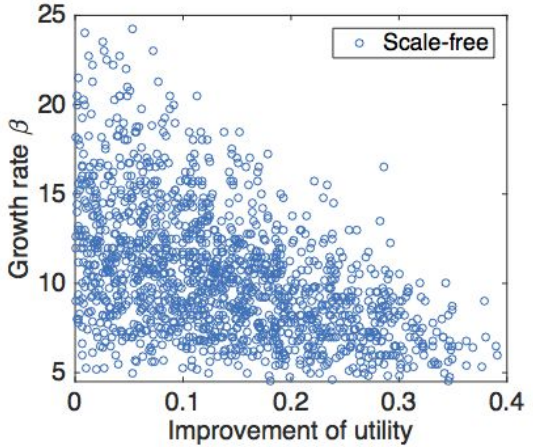
Improvement = (Utility of dynamic flooding - utility of static flooding) / utility of static flooding



(c) Dynamic against static



(d) Random network



(e) Scale-free network

Correlation between beta and the utility improvement on random network is close to zero, indicating that the significance of improvement is irrelevant of a node's growth rate and its position in the network. Meanwhile, such correlation on scale-free network is much stronger, with Pearson correlation being 0.5273.

How Do We Setup the Experiments?

- Let's set up a more realistic experiments.
  - Four realistic ISP networks and a community network.
  - Each node has a 4GB cache with LRU algorithm.
  - Content set is based on a Youtube video trace.
  - Nodes of degree 1 are clients.
  - 10 to 20 servers are randomly selected in a network.
  - The collective request trace is generated using a Hawkes process, which is controlled by both temporal and spatial locality factors.

Do Flooding Strategies Impact Caching?

AS	Byte hit rate			Cost			Avg. hops		
	nw	st	dy	nw	st	dy	nw	st	dy
1239	0.44	0.40	0.43	1.0	0.27	0.28	1.90	1.60	1.62
2914	0.49	0.42	0.47	1.0	0.31	0.32	1.75	1.55	1.58
3356	0.42	0.39	0.42	1.0	0.25	0.27	2.02	1.69	1.74
7018	0.47	0.41	0.45	1.0	0.26	0.28	1.87	1.54	1.63
Guifi	0.51	0.44	0.49	1.0	0.22	0.23	1.71	1.32	1.38

nw: network-wide flooding; **st**: static flooding; **dy**: dynamic flooding.

Network-wide flooding always achieves the best byte hit rate, the improvement is marginal at the price of 2 to 3 times increase cost.

Dynamic flooding consistently outperforms static one.

Most content are discovered within 2 hops. Network-wide flooding has the worst values due to its inherent aggressiveness.

Does Spatial Locality Matter?

- Spatial locality does not play a significant role, especially when content availability is not given as a priori.
  - Higher values improve the hit rate marginally.
  - No impact on cost at all because cost is a function of content and topology, neither will be changed by spatial locality.
- Intuitive explanation: nodes are mostly constrained within a small neighbourhood, and flooding do not go any further into the network. Therefore what is happening outside is not important at all.

What Are the Limitations of This Model?

- Clustering coefficient is not considered in the network model, so it may overestimate the neighbourhood growth.
- Cost of retrieving a content is not considered.
- Sublinear growth in gain and exponential growth in cost, this needs to be verified and justified in reality.
- Only evaluated with LRU, we do not know whether other in-network caching algorithms will change our story or not.

What Are the Takeaways?

- If you cannot get most benefits from nearby neighbours, there is no need to go further in a network.
- The neighbourhood (of a medium scope) can be very well approximated with a node's 2-hop information.
- The choice on static or dynamic flooding depends on the network structure. I.e., random or scale-free networks.
- The results justify the rationale of deploying collaborative caches at network edge from content discovery perspective.

Thank you. Questions?