Content Growth and Attention Contagion in Information Networks:

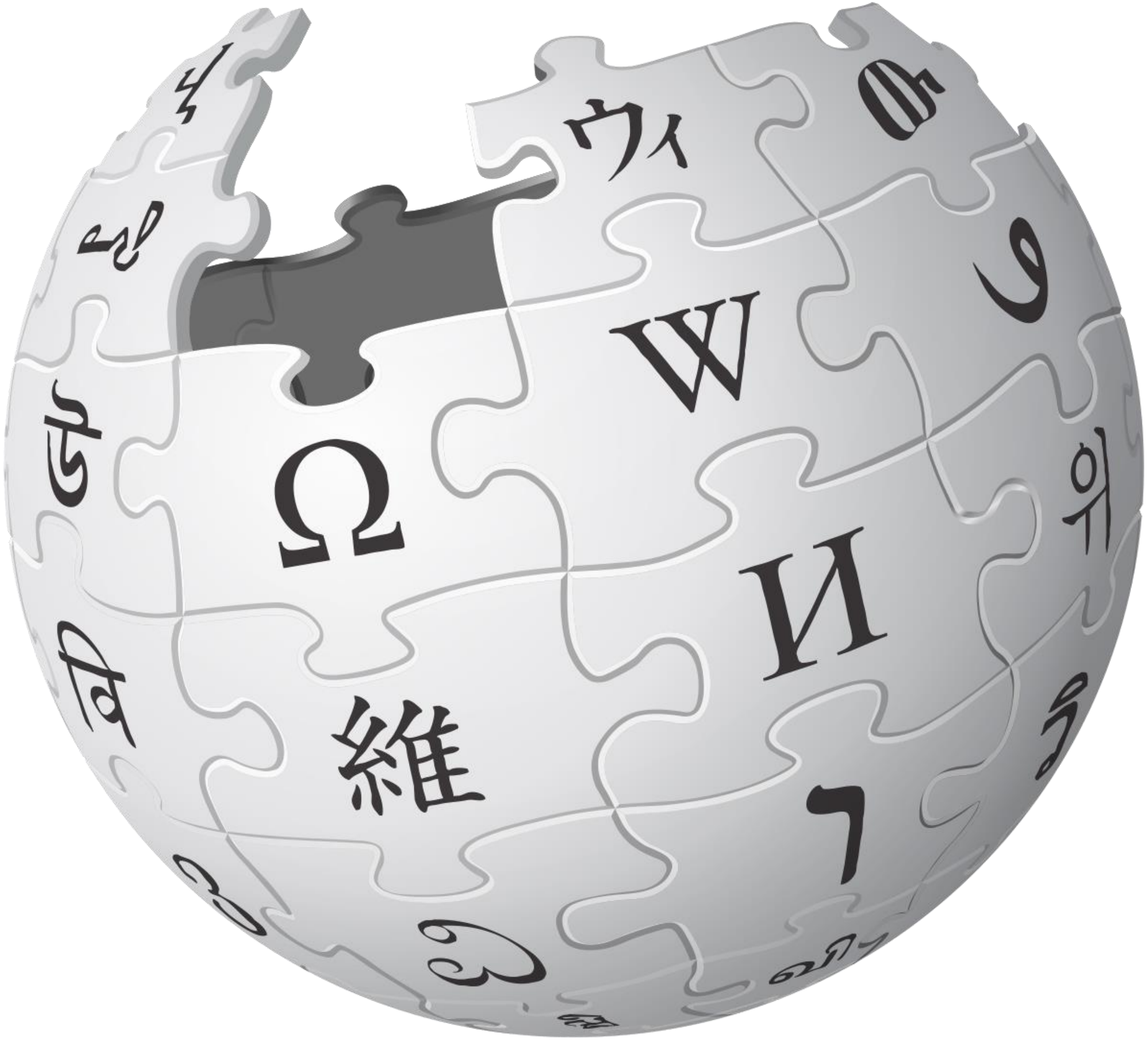
Addressing Information Poverty on Wikipedia



Kai Zhu[1] Dylan Walker[1] Lev Muchnik[2]

1 Boston University Questrom School of Business

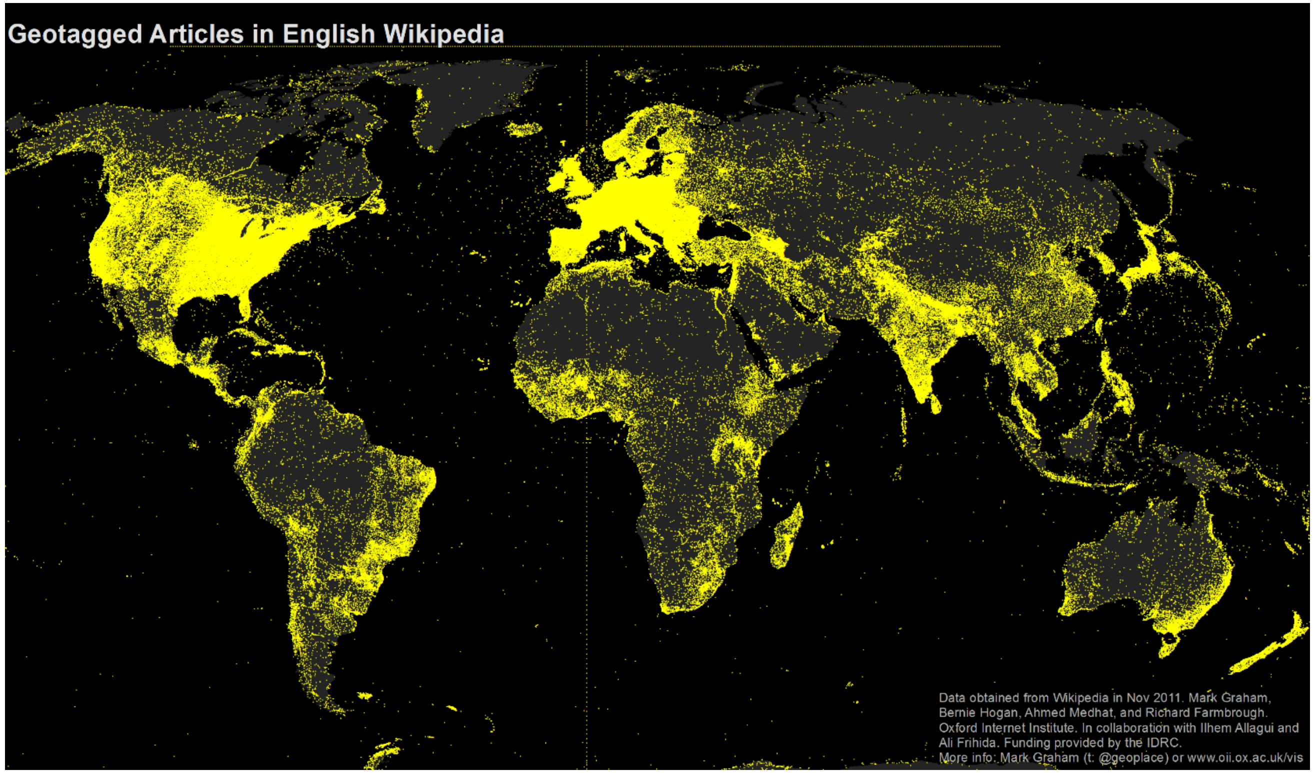
2 THE HEBREW UNIVERSITY OF JERUSALEM



Knowledge Disparities and Information Poverty on Wikipedia

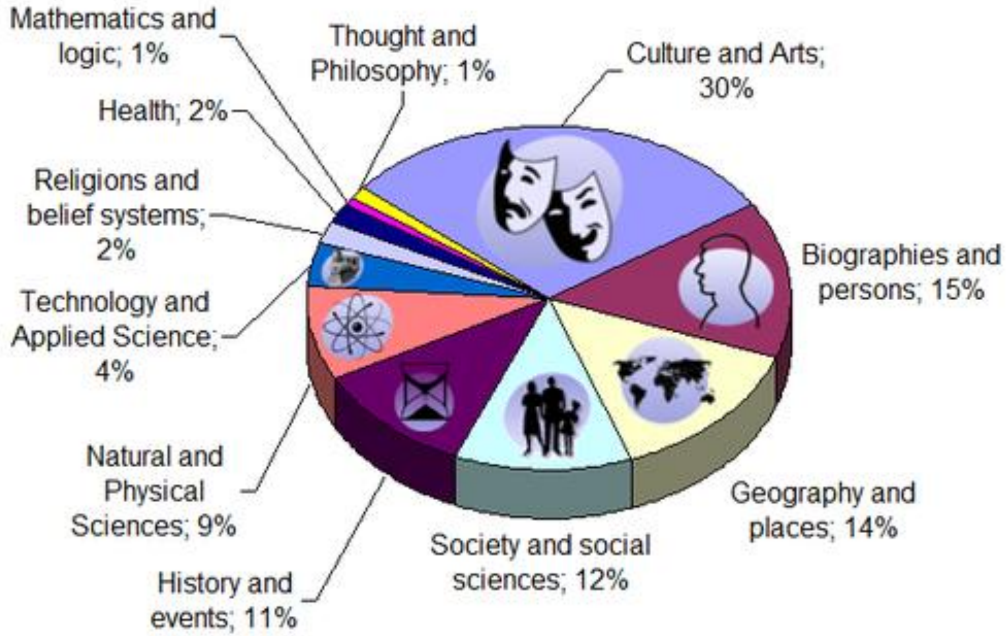
(coverage and depth of knowledge in Wikipedia articles)

Across geographical areas (Graham et al., 2014)



"…some parts of the world remain well below their expected values."

Across knowledge domains (Halavais and Lackaff, 2008)



It's not just Wikipedia – All collaborative information networks are susceptible to the digital divide!

Why? What do we know? What can we do about it?

What do we know?

- Geographic information skews and **digital divides** limit our understanding of and attention to impoverished areas in terms of economic, social, political and cultural concerns (Forman et al. 2012; Norris 2001; Castells 1999; Yu 2006)
  - "Most of Africa is being left in a technological apartheid" – Castells, 1999

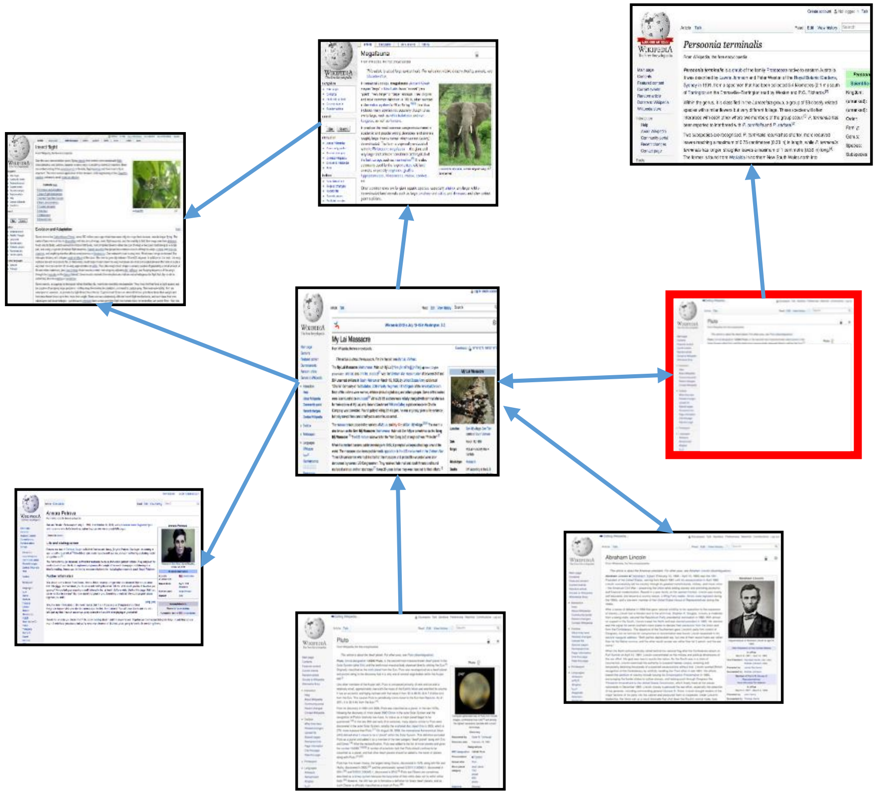
- Info (un)availability has a **strong impact on real-world outcomes** in financial markets, scientific advancement, tourism (Hinnosaur et al. 2017; Thompson and Hanley 2017; Xiaoquan and Lihong 2015; Xu and Zhang 2013)

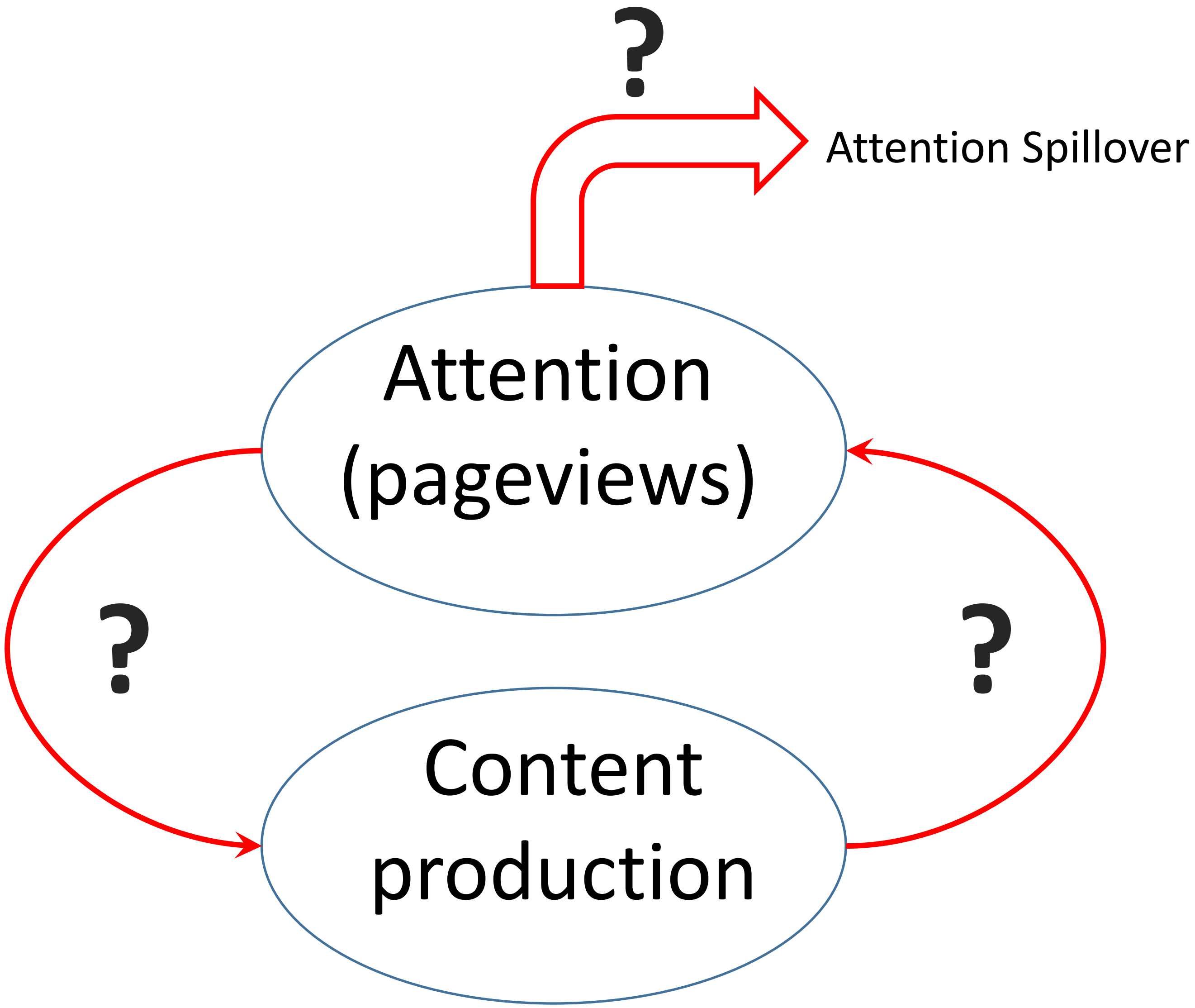
- **Herding and popularity effects** in information networks (Salganik et al. 2006; Muchnik et al. 2013; many others)

What do we know?

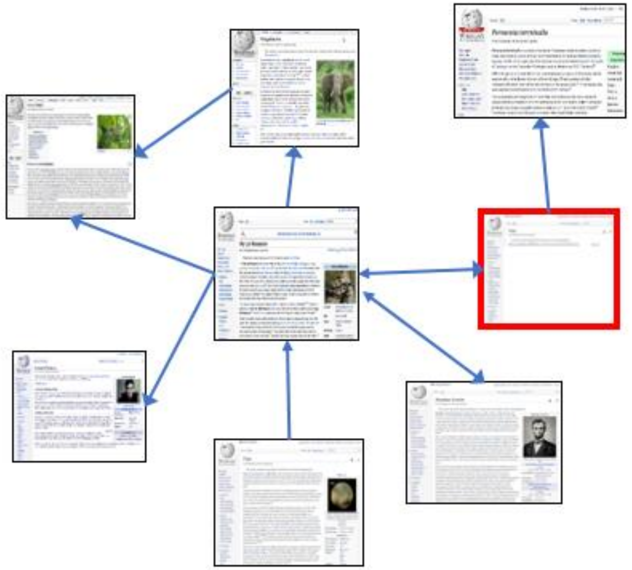
- On Wikipedia:
  - A lot of work on the motivation of editors (Gallus 2016; Lampe et al. 2012; Zhang and Zhu 2011)
  - Production and consumption interact in complex, dynamical ways (Kampf et al. 2012,2015; Wilkinson and Huberman 2007), including **“rich-get-richer” dynamics** (Aaltonen and Seiler 2016)
  - Kane and Ransbotham (2016) find evidence (not causally identified) consistent with a **consumption and production feedback loop**.
  - The network position of an article is correlated with both its production and consumption (Kane, 2009; Kummer et al. 2016; Ransbotham et al. 2012)
  - Structural embeddedness of an article in the content-contributor network is positively correlated with consumption and quality (Kane and Ransbotham 2016; Ransbotham et al. 2012)
  - Still ***no causal estimates*** of the interaction between production and consumption.
- Demand shocks generate **attention that can spillover** on product networks (e.g., Amazon) yielding benefits to downstream products (Carmi et al. 2017)
- Seeding strategies (policies) can leverage spillover in social networks (Aral et al. 2013)

The Causal Identification Challenge





Natural experiment: content shock in Wikipedia

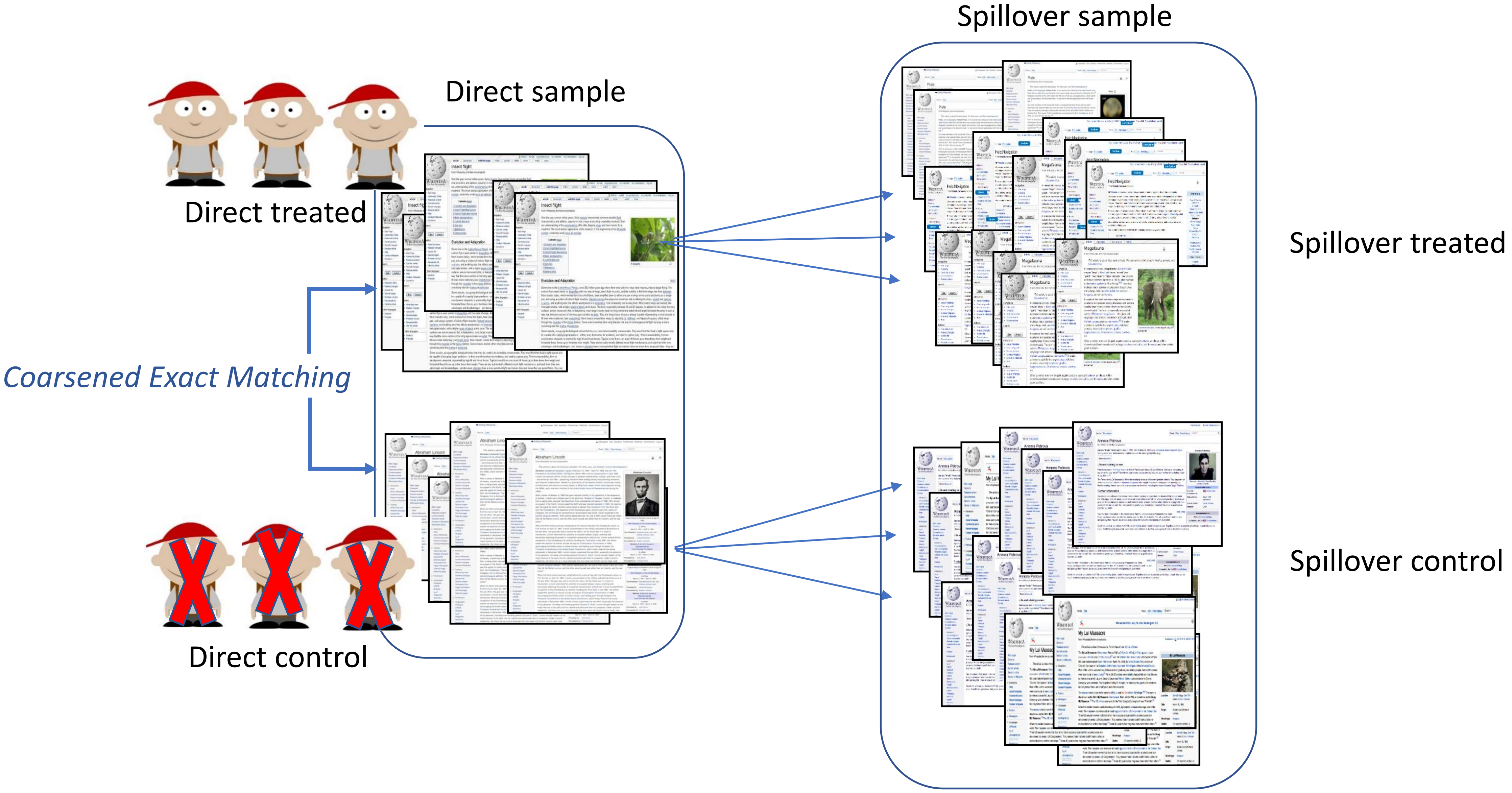


- Enacted through a campaign by the Wikipedia Education Foundation, college students were assigned to expand Wikipedia articles as their class assignments.

- ~35,000 articles expanded or created and ~35M words added to Wikipedia by more than 17,000 students – equivalent to 22 volumes of printed encyclopedia

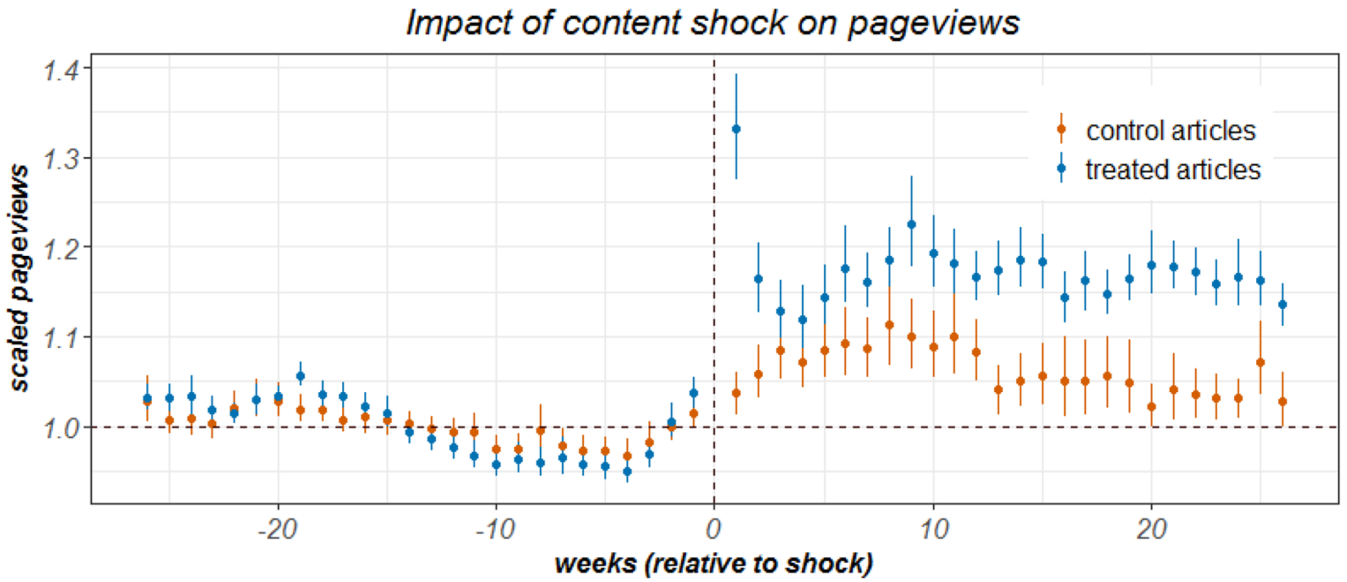
- ***Identification assumption****:* the content contribution by the college students is exogenous to the natural evolution of the articles in the sense that it would not have occurred ***during the same time period*** in the absence of the campaign.

Research Design



Direct Impact of the Content Shock

$$Pageviews_{it} = \beta_1 PostShock_{it} + \beta_2 PostShock_{it} * X_i + \gamma_i + \delta_t + e$$



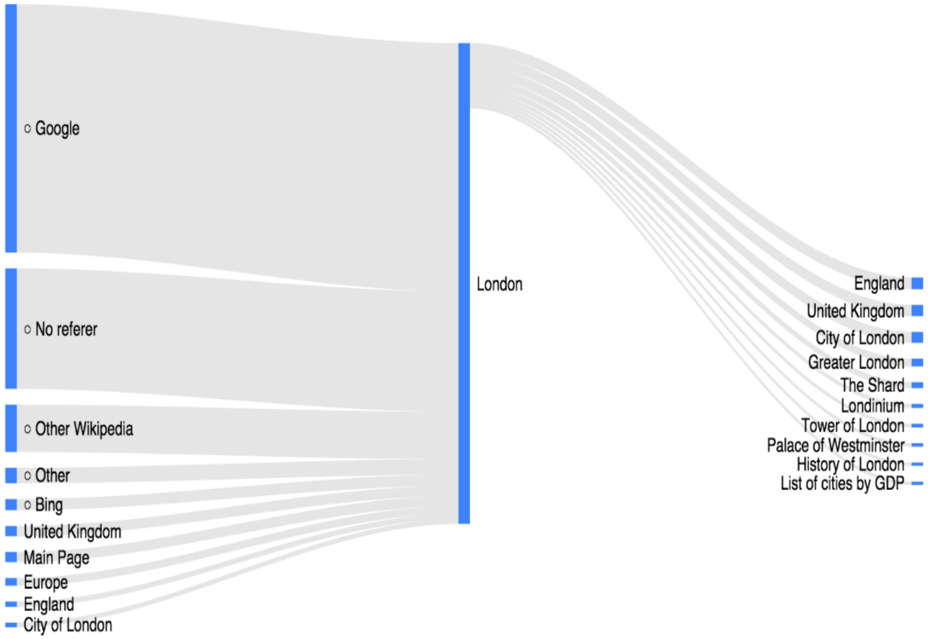
	Scaled pageviews		
	(1)	(2)	(3)
PostShock	0.119***		
	(0.017)		
PostShock*log(char count)		0.035***	0.065***
		(0.005)	(0.008)
PostShock*old article			-0.041*
			(0.024)
PostShock*popular article			-0.142***
			(0.025)
PostShock*long article			-0.015
			(0.025)
Article fixed effect	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes
Observations	287,664	287,664	287,664
Adjusted R^2	0.122	0.122	0.124

- 12% lift in post-shock pageviews on average
- The effect is relatively long-lasting (26 weeks post shock)
- Stronger impact (as high as 30% lift) for less popular articles, with more characters added
- Treated articles received 3.7 more edits (p<1e-9) and 2.2 more unique editors (p<1e-16) on average in the 6 months following the shock [DID estimators; panel models not shown here].

What drives increased attention? i.e., Where does it come from?

Clickstream Data and the Sources of Increased Attention

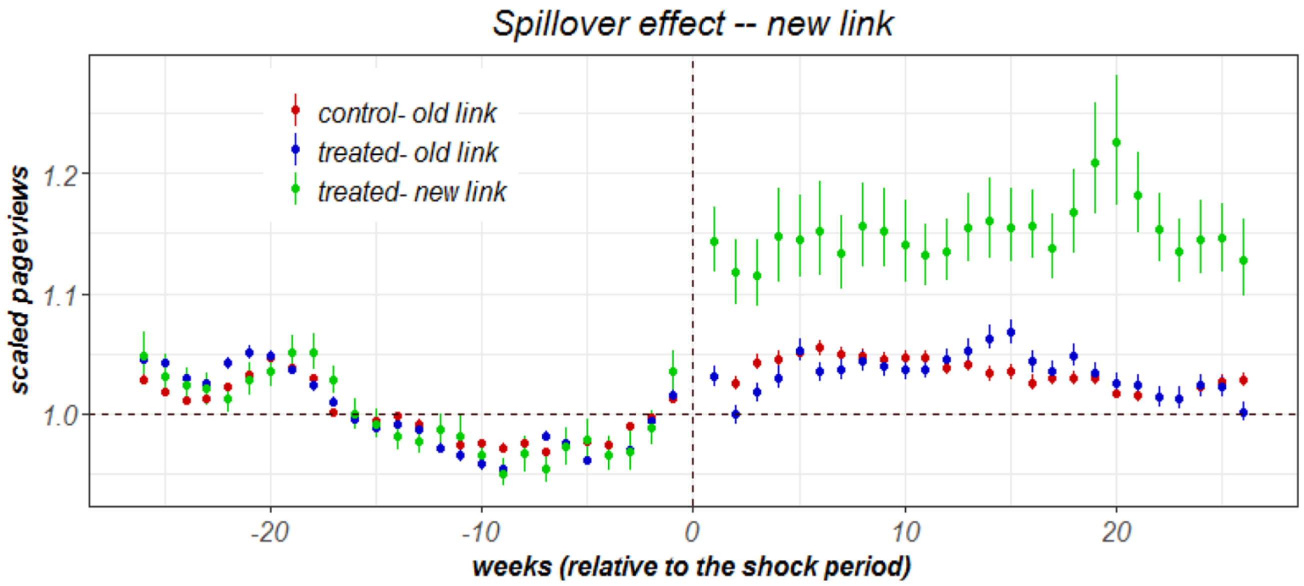
- Wikimedia has recently made available monthly clickstream data that details the cumulative web traffic to each article from external sources and across internal links (from one article to another)
  - ~ 26M (referrer, resource) pairs over ~ 8B web requests



Wikipedia Clickstream dataset

- We use this data to compare traffic sources for treated vs. control articles
- We find increased traffic comes from both internal links and external websites
  - internal traffic is explained by more incoming links (0.5 on average) added for treated articles
  - external traffic is explained by improved search engine visibility (4.8 more visits/day)

Spillover of Attention



	Scaled pageviews			
	(1)	(2)	(3)	(4)
PostShock	0.008***	0.027***	-0.006	-0.005
	(0.003)	(0.006)	(0.004)	(0.007)
PostShock*popularTargetArticle		-0.013**		-0.004
		(0.005)		(0.005)
PostShock*popularSourceArticle		-0.016**		0.000
		(0.007)		(0.007)
PostShock*newLink			0.129***	0.148***
			(0.012)	(0.018)
PostShock*popularTargetArticle*newLink				-0.138***
				(0.023)
PostShock*popularSourceArticle*newLink				0.073***
				(0.023)
Article fixed effect	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes
Observations	6,862,648	6,862,648	6,862,648	6,862,648
Adjusted R^2	0.104	0.104	0.104	0.104

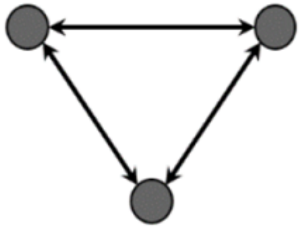
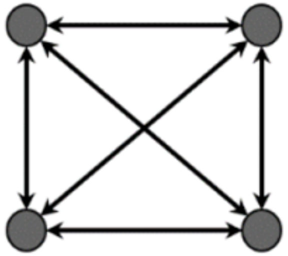
- A new link brings significant traffic on average (12.9% lift)
- Stronger impact (as high as 22.1%) for new links from popular source articles to unpopular target articles; and (14.8% lift) when both source, target are unpopular.
- This suggests that articles in impoverished regions may stand to benefit substantially from spillover.

Policy to Harness Attention Contagion

- Our causal estimates show that attention to articles spills over onto downstream articles and this is effect is most prevalent over newly created links and to less popular downstream articles.
  - In social networks, we might term this spillover as contagious – articles are more likely to “catch” attention from upstream articles.
  - Can we create a policy to leverage this effect to benefit **information-impoverished regions** in the network?
- Attention Contagion Policy (ACP): Encourage editors to focus their efforts on highly related (connected) topics/groups of articles. In network terms, this means focusing on cliques or communities of articles.
  - How might we measure this?
  - We need a baseline to compare it to....
- Undirected Attention Policy (UAP): Editors focus their attention on articles without considering their relatedness or network structure.

Intuition: a mean-field estimation of spillover benefit

- Assume all articles get (the same) direct traffic $\boldsymbol{T}$
- Assume surfers have an equal tendency to follow any link
- Assume (identical) spillover across a single link is $\boldsymbol{fT}$
- Assume no backtracking or repeat visits
- For now, look only at cliques (same benefit for all articles)

Clique				…	n-clique
Total Spillover	$2f$	$6f + 6f^2$	$12f + 24f^2 + 24f^3$	…	$\sum_{k=2}^{n} \underbrace{\frac{n!}{(n-k)!} f^{k-1}}$

All partial permutations of a set of k articles, describing a path of length $(k-1)$ that successive spillover take.

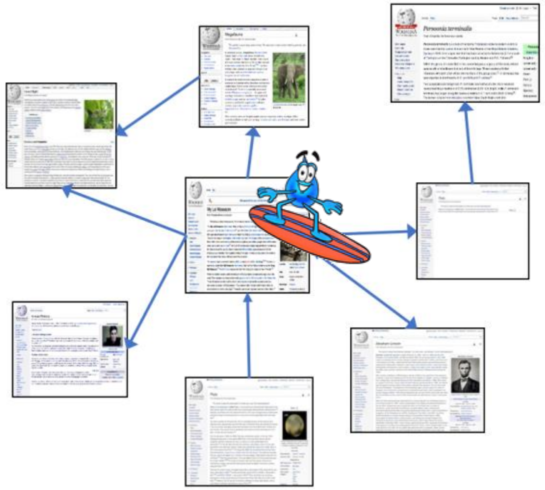
- Demonstrates the intuition of capturing the benefit of Attention Contagion Policies
- Permits direct estimation of the benefit under mean-field assumptions.
- But the real-world doesn't obey the above assumptions. We can do better!

Empirically-Informed Diffusion Simulation

- We start with the well-known diffusion model PageRank:

$$\vec{r}_{t+1} = (1-\alpha)\vec{r}_0 + \alpha G \cdot \vec{r}_t$$

  - At each time step, random surfers hop to a new article with probability $(1-\alpha)$ or instead (with probability α) follow a randomly chosen link to another article.
  - Keep doing this until convergence: $|\vec{r}_{t+1} - \vec{r}_t| < \epsilon$, which defines the PageRank number $\vec{r}$

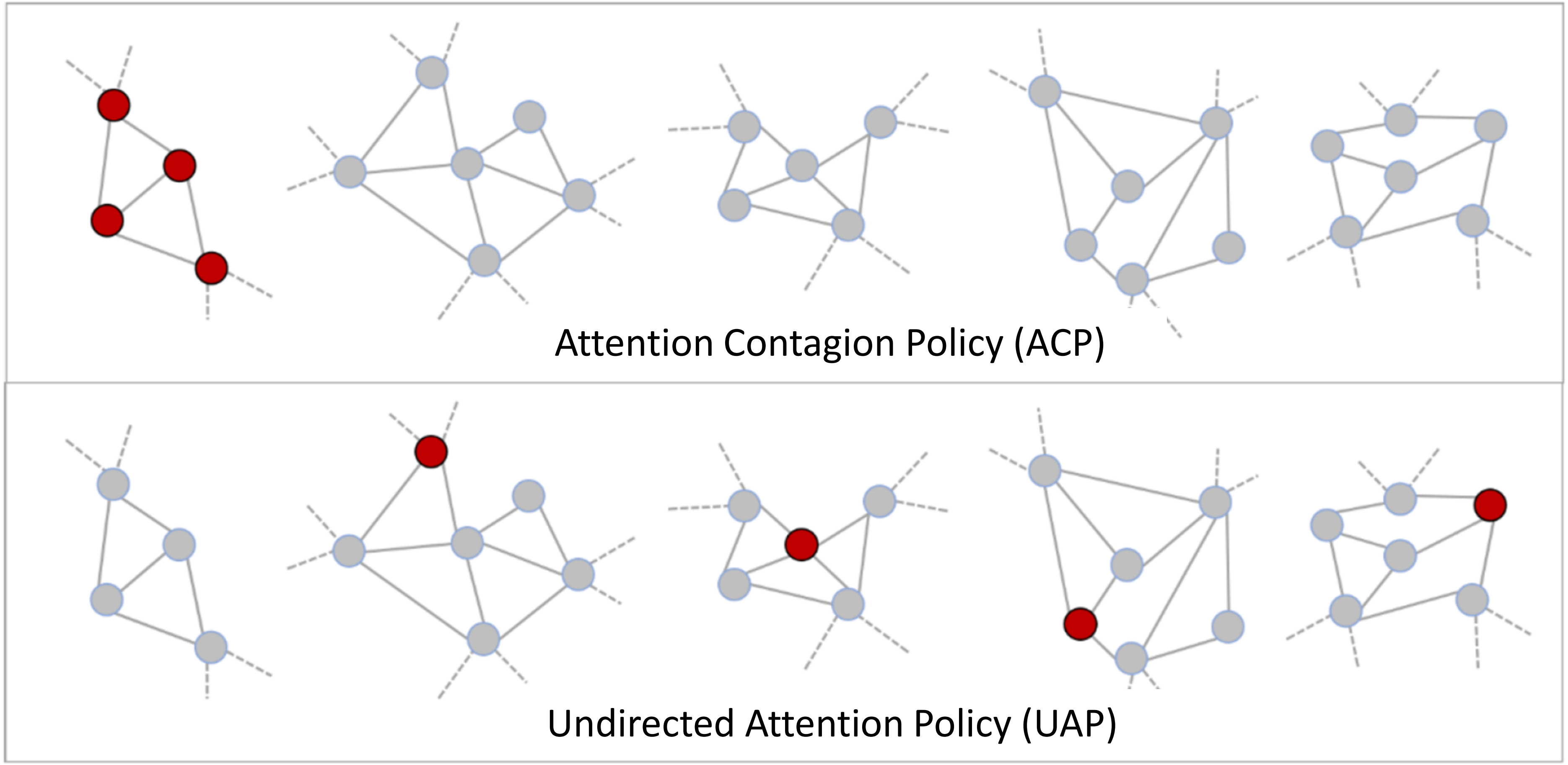


- Vanilla PageRank:
  - $G_{ij} = A_{ij}/k_j$ (assumes random choice of link to follow)
  - $\vec{r}_0 = \vec{1}$ (assumes equal chance of landing on any article when hopping)
- We can do better and make diffusion follow empirical data on actual surfing behavior by using **clickstream data**:
  - $G_{ij} \sim$ empirical probability to follow a link
  - $\vec{r}_0 \sim$ empirical probability to land on an article from an external source
  - This allows us to obtain a steady state for traffic: $\vec{r} = PR(\vec{r}_0, G, \alpha, \epsilon)$
- So what happens when some articles get a **shock to attention**?
  - We can simulate this by perturbing incident external traffic (for some set of articles) and measuring its impact on attention to all articles in the system:

$$\boldsymbol{\vec{r}_p^S = PR(\vec{r}_{0p}^S, G, \alpha, \epsilon)}$$

A Tale of Two Policies

- We need a way to find “impoverished regions”:
  - Search for maximal cliques and communities in the weighted network (from clickstream traffic data) with “low traffic
- So here’s what a single perturbative simulation would look like:



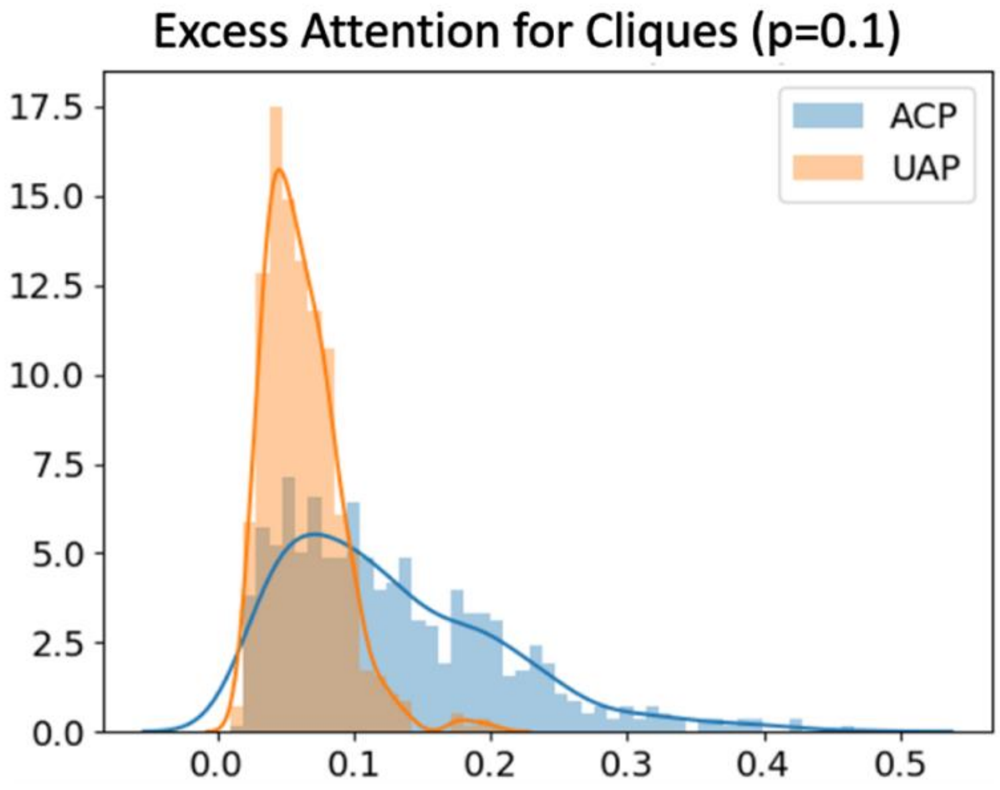
For a given simulation of ACP , we select a community (or clique) and perturb all the nodes (red) in that subnetwork.
We create a matching simulation of UAP, where each node in ACP is matched to a node in UAP.

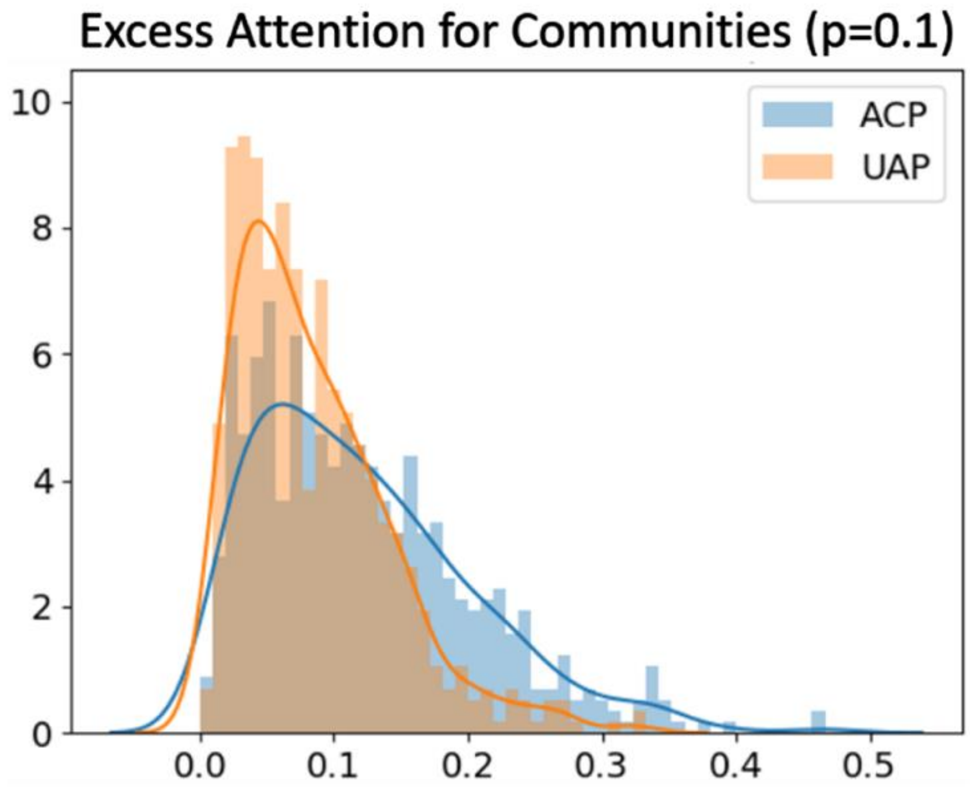
Excess Attention – Benefits of Spillover

Excess Attention: $EA(S,p) = \sum_{i\in S} \frac{r_{p,i}^{S} - r_i}{r_i}$

The percentage difference in PageRank number (relative to no perturbation) for articles in the perturbed set (S)

We choose 600 cliques/communities with "low traffic" (impoverished) and perform this simulation:





The Attention Contagion Policy leads to significant increases (up to 2x on average) in Excess Attention (p<1e-71).

It harness attention contagion to benefit impoverished regions in the information network.

Takeaway

- Directed editorial efforts to develop underdeveloped articles have significant and long-lasting impact
  - A positive *feedback loop* between content production and consumption in open collaboration systems.
- Attention propagates over the information network through *hyperlinks*
  - Attention spillover is particularly strong for new links and less popular linked articles
- Informational inequities can be alleviated using policies that best leverage *attention spillovers*