Online Social Networks and Media

Fairness, Diversity

Outline

- Fairness (case studies, basic definitions)
- Diversity
- An experiment on the diversity of Facebook

Fairness, Non-discrimination

To **discriminate** is to treat someone differently

(Unfair) discrimination is based on *group membership*, *not individual merit*

Some attributes should be irrelevant (protected)

Disparate treatment and impact

Disparate treatment: Treatment depends on class membership

Disparate impact: Outcome depends on class membership
(Even if (apparently) people are treated the same way)

Doctrine solidified in the US after [Griggs v. Duke Power Co. 1971] where a high school diploma was required for unskilled work, excluding black applicants

Case Study: Gender bias in image search [CHI15]

What images do people choose to represent careers?

In search results:

- evidence for *stereotype exaggeration*
- systematic *underrepresentation of women*

- People rate search results *higher* when they are *consistent* with stereotypes for a career
- Shifting the representation of gender in image search results can *shift* *people's perceptions* about real-world distributions. (after search slight increase in their believes)

Tradeoff between high-quality result and broader societal goals for equality of representation

Case Study: Latanya

The importance of being Latanya

Names used predominantly by *black men and women* are much more likely to generate *ads* related *to arrest records*, than names used predominantly by white men and women.

Case Study: AdFisher

Tool to automate the creation of *behavioral* and *demographic* profiles.

http://possibility.cylab.cmu.edu/adfisher/

- setting gender = female results in less ads for high-paying jobs
- browsing substance abuse websites leads to rehab ads

Case Study: Capital One

Capital One uses tracking information provided by the tracking network [x+1] to personalize offers for credit cards

Steering minorities into higher rates

capitalone.com

Fairness: google search and autocomplete

Donald Tramp accused Google “suppressing negative information” about Clinton
Autocomplete feature - “hillary clinton cri” vs “donald tramp cri”

Autocomplete:

- are jews
- are women

https://www.theguardian.com/us-news/2016/sep/29/donald-trump-attacks-biased-lester-holt-and-accuses-google-of-conspiracy

https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook?CMP=fb_gu

Google+ names

Google+ tries to classify Real vs Fake names

Fairness problem:

– Most training examples standard white American names

– Ethnic names often unique, much fewer training examples

Likely outcome:

Prediction accuracy *worse on ethnic names*

Katya Casio. *"Due to Google's ethnocentricity I was prevented from using my real last name (my nationality is: Tungus and Sami)"*
Google Product Forums

Other

LinkedIn: female vs male names (for female prompts suggestions for male, e.g., Andrea Jones" to "Andrew Jones," Danielle to Daniel, Michaela to Michael and Alexa to Alex.)
http://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

Flickr: auto-tagging system labels images of black people as apes or animals and concentration camps as sport or jungle jyms.
https://www.theguardian.com/technology/2015/may/20/flickr-complaints-offensive-auto-tagging-photos

Airbnb: race discrimination
Against guest
http://www.debiasyourself.org/
Community commitment
http://blog.airbnb.com/the-airbnb-community-commitment/
Non-black hosts can charge ~12% more than black hosts
Edelman, Benjamin G. and Luca, Michael, Digital Discrimination: The Case of Airbnb.com (January 10, 2014). Harvard Business School NOM Unit Working Paper No. 14-054.

Google maps: China is about 21% larger by pixels when shown in Google Maps for China
Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson: MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. Proc. of WWW. Montreal, Quebec, Canada, April 2016

Reasons for bias/lack of fairness

Data input

- Data as a social mirror: Protected attributes redundantly encoded in observables
- Correctness and completeness: Garbage in, garbage out (GIGO)
- Sample size disparity: learn on majority (Errors concentrated in the minority class)
- Poorly selected, incomplete, incorrect, or outdated
- Selected with bias
- Perpetuating and promoting historical biases

Reasons for bias/lack of fairness

Algorithmic processing

- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Fairness through blindness

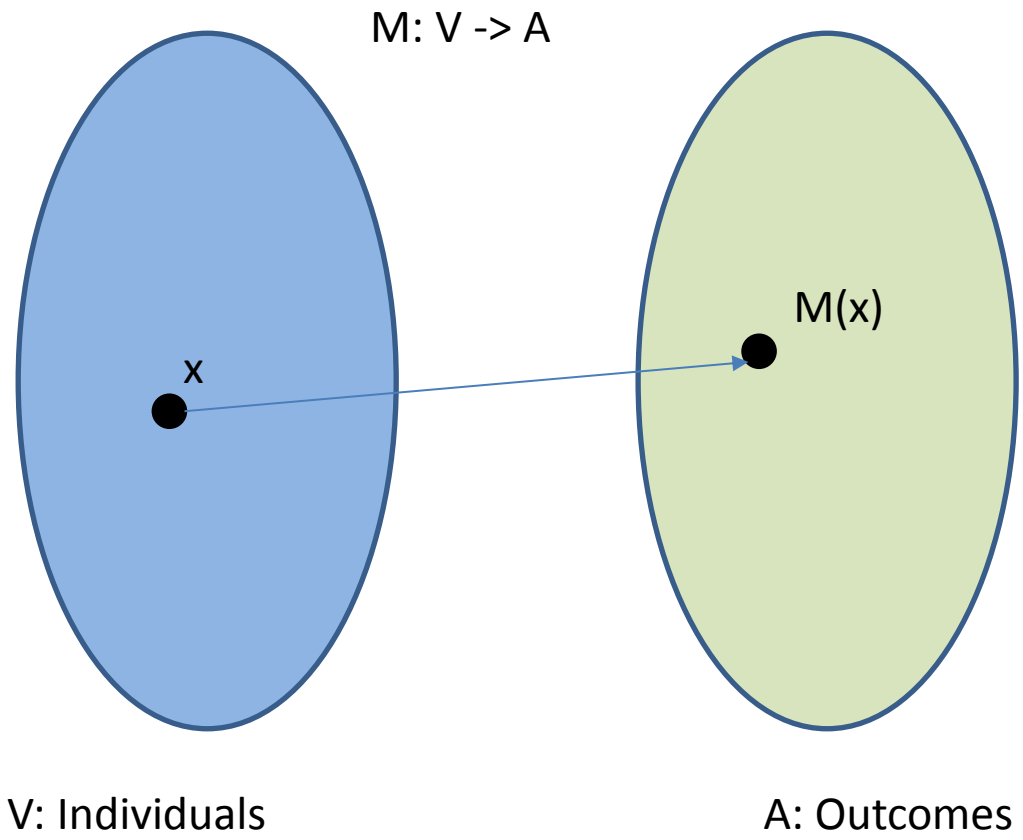
Ignore all irrelevant/protected attributes

Useful to avoid formal disparate treatment

Fairness: definition

- Classification
  - Classification/prediction for people with similar non-protected attributes should be similar
  - Differences should be mostly explainable by non-protected attributes

  - A (trusted) data owner that holds the data of individuals, a *vendor* that classifies the individuals


M: V -> A
x
M(x)
V: Individuals
A: Outcomes

Main points

- Individual-based fairness: any two individuals who are *similar* with respect to a particular task should be *classified similarly*
- Optimization problem: construct fair classifiers that minimize the expected *utility loss* of the vendor

Formulation

V: set of individuals

A: set of classifier outcomes
Classifier maps individuals to outcomes

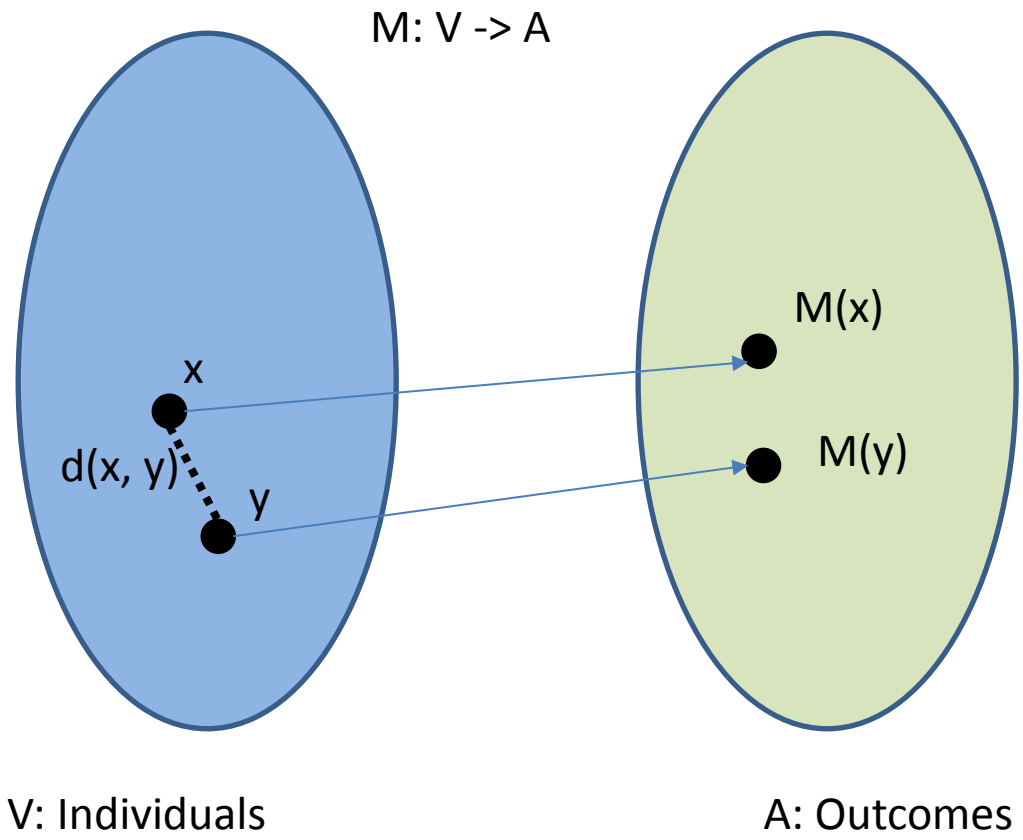
Randomized *mapping* $M: V \to \Delta(A)$ from individuals to probability distributions over outcomes

- To classify $x \in V$, choose an outcome a according to distribution $M(x)$

Formulation

A task-specific distance metric *d*: *V* x *V* -> R on individuals

- Expresses *ground truth* (or, best available approximation)
- Public
- Open to discussion and refinement
  - Externally imposed, e.g., by a regulatory body, or externally proposed, e.g., by a civil rights organization


M: V -> A
x
d(x, y)
y
M(x)
M(y)
V: Individuals
A: Outcomes

Formulation

Lipschtiz Mapping: a mapping M: V -> Δ(A) satisfies the (D, d)-Lipschitz property, if for every x, y ∈ V, it holds

$$D(M(x), M(y)) \leq d(x, y)$$

Formulation

There exists a classifier that satisfies the Lipschitz condition

- Map all individuals to the same distribution over outcomes

Vendors specify arbitrary utility function

$$U: V \times A \rightarrow R$$

Find a mapping from individuals to distributions over outcomes that *minimizes expected loss* subject to the Lipschitz condition.

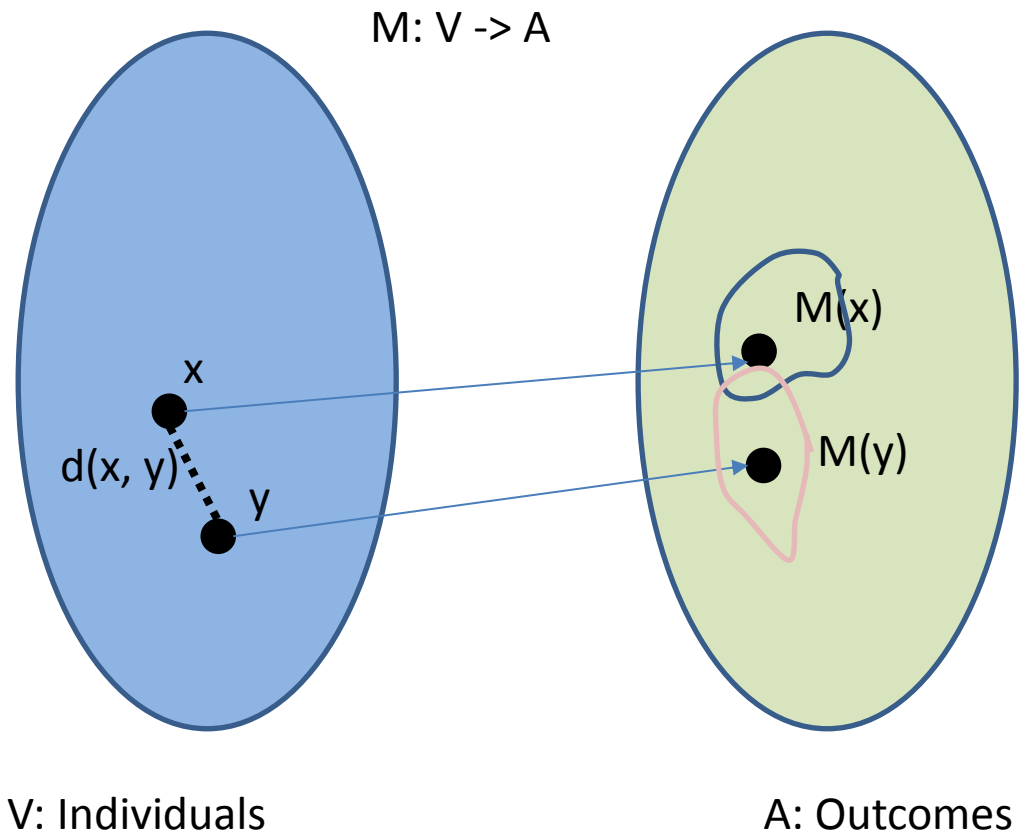
Formulation

$$\mathrm{opt}(\mathcal{I}) \stackrel{\mathrm{def}}{=} \min_{\{\mu_x\}_{x \in V}} \mathop{\mathbf{E}}_{x \sim V} \mathop{\mathbf{E}}_{a \sim \mu_x} L(x, a) \tag{2}$$

$$\text{subject to} \quad \forall x, y \in V, : \quad D(\mu_x, \mu_y) \leq d(x, y) \tag{3}$$

$$\forall x \in V : \quad \mu_x \in \Delta(A) \tag{4}$$

What is *D*?



What is *D*?

Statistical distance or local variation between two probability measures *P* and *Q* on a finite domain *A*

$$D_{lv} = \frac{1}{2}\sum_{a \in A} |P(a) - Q(a)|$$

Example
A = {0, 1}

Most different
P(0) = 1, P(1) = 0
Q(0) = 0, Q(1) = 1
D(P, Q) = 1

Most similar
P(0) = 1, P(1) = 0
Q(0) = 1, Q(1) = 0
D(P, Q) = 0

P(0) = P(1) = 1/2
Q(0) = 1/4, Q(1) = 3/4
D(P, Q) = 1/4

Assumes d(x, y) close to 0 for similar and close to 1 for dissimilar

What is *D?*

$$D_\infty(P,Q)$$

$$= sup_{a \in A} log\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right)$$

Example

A = {0, 1}

Most different

P(0) = 1, P(1) = 0

Q(0) = 0, Q(1) = 1

Most similar

P(0) = 1, P(1) = 0

Q(0) = 1, Q(1) = 0

P(0) = P(1) = 1/2

Q(0) = 1/4, Q(1) = 3/4

Statistical parity (group fairness)

If M satisfies statistical parity, then members of S are equally likely to observe a set of outcomes O as are not members

$$|\Pr\{M(x) \in O \mid x \in S\} - \Pr\{M(x) \in O \mid x \in S^c\}| \leq \varepsilon$$

If M satisfies statistical parity, the fact that an individual observed a particular outcome provides no information as to whether the individual is a member of S or not

$$|\Pr\{x \in S \mid M(x) \in O\} - \Pr\{x \in S^c \mid M(x) \in O\}| \leq \varepsilon$$

Catalog of evils

1. Blatant explicit discrimination:

membership in *S* *explicitly tested* for and *a worse outcome* is given to members of *S* than to members of S^c

2. Discrimination Based on Redundant Encoding:

Explicit test for membership in *S* *replaced by an essentially equivalent test*

successful attack against "fairness through blindness"

Catalog of evils

3. Redlining:

well-known form of discrimination based on redundant encoding.

Definition [Hun05]: "the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor."

4. Cutting off business with a segment of the population in which membership in the protected set is disproportionately high:

generalization of redlining, in which members of *S* need not be a majority; instead, the fraction of the redlined population belonging to *S* may simply *exceed* the fraction of *S* in the population as a whole.

Catalog of evils

5. Self-fullfilling prophecy:

Deliberately *choosing the "wrong" members of S* in order *to build a bad "track record"* for S

A less malicious vendor simply selects *random members of S* rather than qualified members

6. Reverse tokenism:

Goal is to create convincing refutations

Deny access to a qualified member of S^c

c is a token rejectee

References

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard S. Zemel: *Fairness through awareness*. ITCS 2012: 214-226

Diversity:
Why, What, How

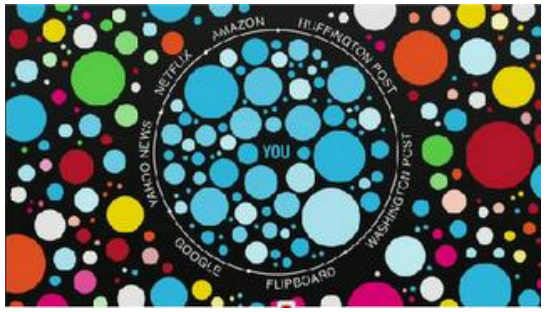
Talk at Dagstuhl seminar on "*Data, Responsibly*", July 2016
With Marina Drosou

Why?

Over Personalization

Filter Bubble: Search results, browsing, recommendations (friends, things, information, …) based on user profiles (own past behavior, similar people, friends, …)

Echo chambers: individuals are exposed only to information from like-minded individuals



What the majority likes

Ranking based on popularity:

popular items get more popular

Other bias

Political, economical, .. (sponsored)

Besides search results diversity also in

Summaries (e.g., reviews) or representatives

Forming committees or teams

Diversity is good

- *No useful information is missed*: results that cover all user intents
- *Better user experience*: less boring, more interesting, human desire for discovery, variety, change
- *Personal growth*: limited, incomplete knowledge, a self-reinforcing cycle of opinion

Better (Fair? Responsible?) decisions

Filter Bubble – Eco Chambers: an experiment

Created two Facebook accounts
"Rusty Smith", *right-wing avatar*, liked a variety of conservative news sources, organizations, and personalities, from the Wall Street Journal and The Hoover Institution to Breitbart News and Bill O'Reilly.
"Natasha Smith", *left-wing avatar*, liked The New York Times, Mother Jones, Democracy Now and Think Progress.

Ten US voters – five conservative and five liberal – liberals were given log-ins to the conservative feed, and vice versa

https://www.theguardian.com/us-news/2016/nov/16/facebook-bias-bubble-us-election-conservative-liberal-news-feed

What?

Aspects of diversity (varying in their relevance to fairness)

The Data Diversity Problem

Given a set P of n items

Select a subset $\boldsymbol{S} \subseteq P$ *with the most diverse items* in P

Variations of the problem:

- *(size) Top-k*: the ***k* most diverse** items in P
- *(quality) Threshold*: items with diversity larger than some threshold value

Coverage

Assuming different topics (e.g., concepts, categories, aspects, intents, interpretations, perspectives, opinions, etc)
Find items that cover all (most) of the topics

For example,
Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, Samuel Ieong: *Diversifying search results*. WSDM 2009

We get the “car” and the “animal” topics but also a “team”, a “guitar”, etc ..

- Assumes “known” topics

jaguar

All Images Videos News Maps More Search tools

About 99,000,000 results (0.46 seconds)

Jaguar
www.jaguar.com/market-selector.html
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets and model-specific websites.

Luxury saloons, sports cars & performance SUV | Jaguar Cars
www.jaguar.com/index.html
The official website of Jaguar. Discover the Art of Performance with our range of vehicles from the XE, XF, XJ, F-Type and, our luxury SUV, the F-PACE.

Images for jaguar Report images

More images for jaguar

Jaguar Cars - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Jaguar_Cars
Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England, owned by ...

Jaguar - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Jaguar
The jaguar (Panthera onca) is a big cat, a feline in the Panthera genus, and is the only extant Panthera species native to the Americas. The jaguar is the ...

Αρχική | Jaguar Greece
www.jaguar.gr/ Translate this page
το αλοκαινουργο PREMIUM μεσαιου μεγεθους SPORTS SALOON της jaguar. ... Το συναρπαστικό νέο SUV επιδόσεων της Jaguar συνδυάζει τη σπορ ...

Jaguar - Facebook
https://www.facebook.com/Jaguar/
Jaguar. 14962244 likes · 7640 talking about this · 116 were here. Jaguar. The Art of Performance.

Jaguar | Basic Facts About Jaguars | Defenders of Wildlife
www.defenders.org/jaguar/basic-facts
The jaguar is the largest cat in the Americas. The jaguar has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color ...

Jaguar (@Jaguar) | Twitter
https://twitter.com/jaguar?lang=en
6373 tweets • 1752 photos/videos • 773K followers. Check out the latest Tweets from Jaguar (@Jaguar)

English | Jaguar Land Rover Corporate Website
www.jaguarlandrover.com/
English | Latest corporate news, information & reports from Jaguar Land Rover.

Jaguar Cars
Luxury vehicles company

jaguar.com/market-selector.html

Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England, owned by the Indian company Tata Motors since 2008. Wikipedia

CEO: Ralf Speth

Founded: September 4, 1922, Blackpool, United Kingdom

Headquarters: Coventry, United Kingdom

Parent organizations: Jaguar Land Rover, Tata Motors

Founders: William Walmsley, William Lyons

Profiles

Twitter YouTube Facebook LinkedIn

Feedback

See results about

Jaguar (Animal)
Conservation status: Near Threatened (Po..
Scientific name: Panthera onca

Content Dissimilarity

Assuming (multi-dimensional, multi-attribute) items + a *distance measure (metric)* between the items
Find the most *different/distant/dissimilar* items

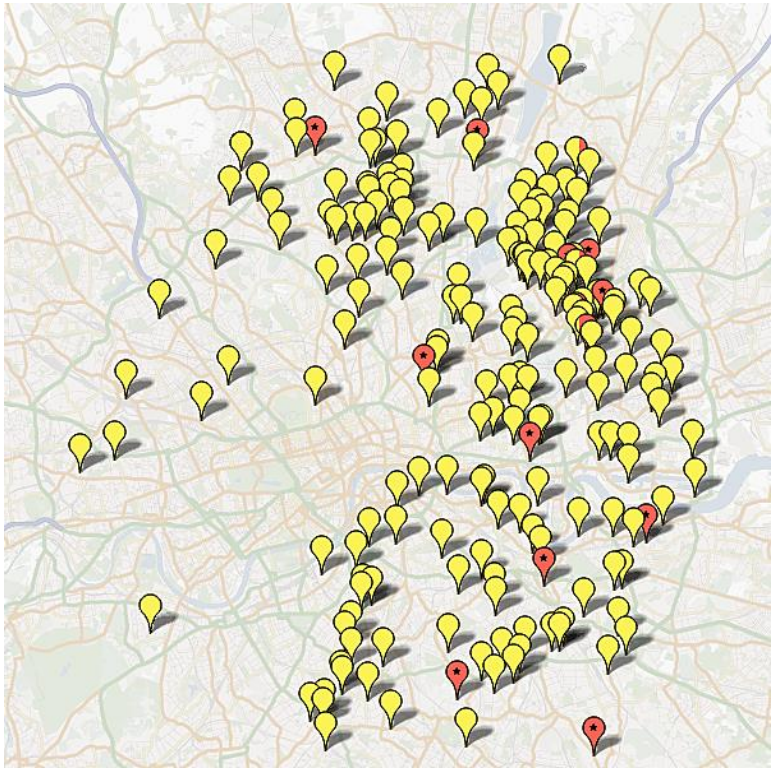
- Distance depends on the items and the problem
- Diversity ordering of the attributes

Defining distance/dissimilarity is key

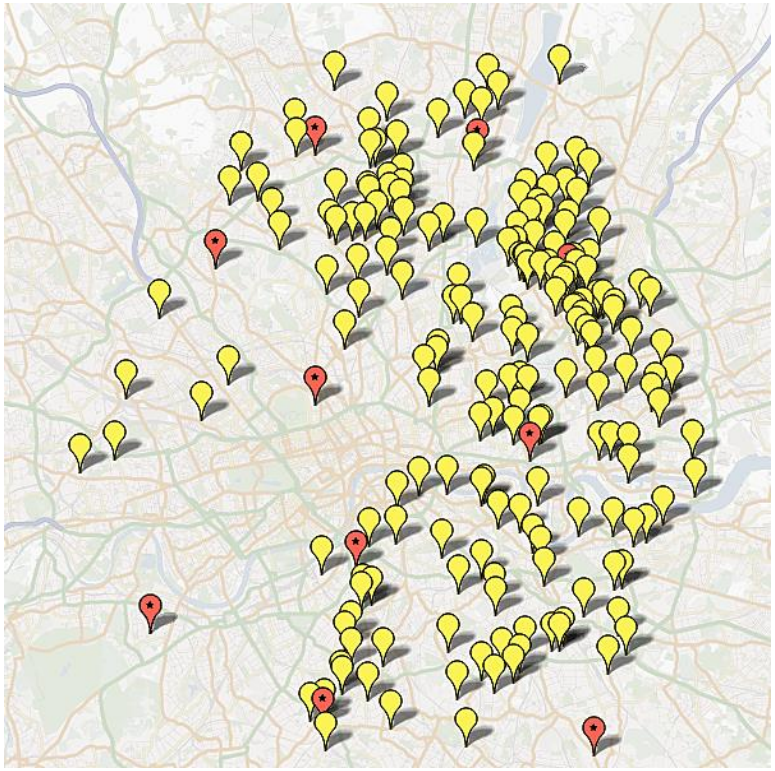
For example, Sreenivas Gollapudi, Aneesh Sharma: *An axiomatic approach for result diversification*. WWW 2009

Example: Two-bedroom apartments up to $300K in London

Top based on price *without* (location) diversity



Top based on price *with* (location) diversity

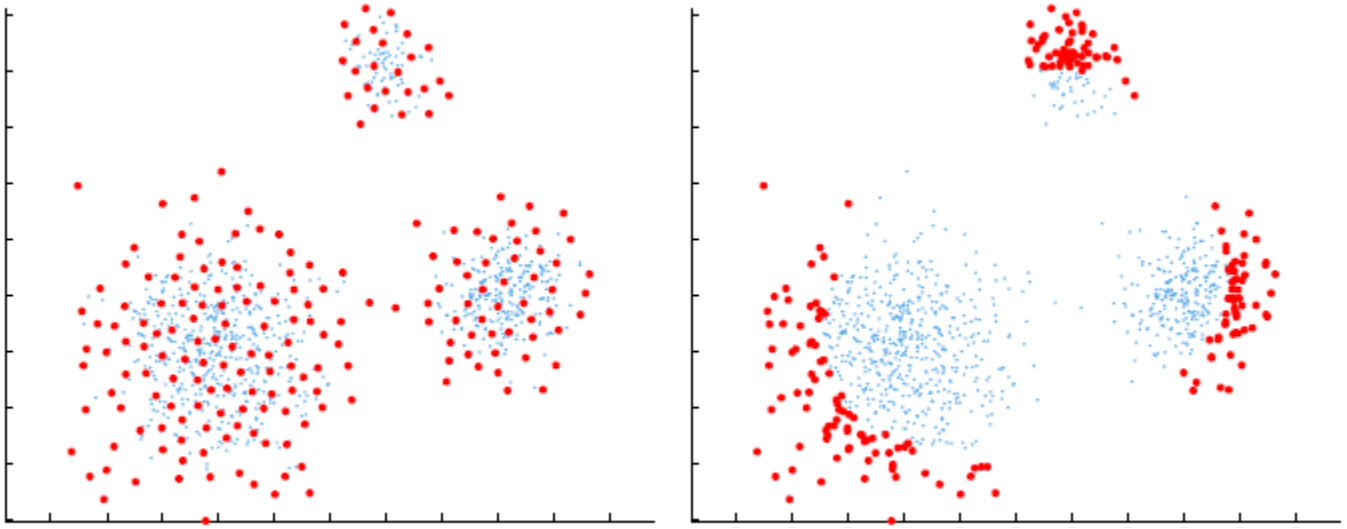


Maximize Set Diversity

Given a distance measure ***d*** and a function ***f*** measuring the diversity of *set* of *k* items,

$$S^* = \underset{\substack{S \subseteq P \\ |S| = k}}{\operatorname{argmax}} \; f(S, d)$$

$$f_{\text{MIN}}(S, d) = \min_{\substack{p_i, p_j \in S \\ p_i \neq p_j}} d(p_i, p_j) \qquad f_{\text{SUM}}(S, d) = \sum_{\substack{p_i, p_j \in S \\ p_i \neq p_j}} d(p_i, p_j)$$



(a) MaxMin. (b) MaxSum.

Novelty

Assuming the *history* of items seen in the past
Find the items that are the *most diverse (coverage, distance)* with respect to what a user (or, a community) *has seen in the past*

- *Marginal relevance*
- *Cascade (evaluation) models:* users are assumed to *scan result lists* from the top down, *eventually stopping* because either their information need is satisfied or their patience is exhausted

Novelty

Relevant concept: serendipity
represents the “unusualness" or “surprise“
(some notion of semantics – the guitar vs the animal)

For example, Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, Ian MacKinnon: *Novelty and diversity in information retrieval evaluation.* SIGIR 2008
Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, Tamas Jambor: *Auralist: introducing serendipity into music recommendation.* WSDM 2012

Multi-criteria

Diversity (coverage, dissimilarity, novelty, serendipity) is just *one of the criteria* in data selection or ranking

E.g., relevance in IR or accuracy in recommendations

MaxSum diversification: maximize the sum (average) relevance (r) and dissimilarity

$$score(S) = (k-1)\sum_{u \in S} r(u) + 2\lambda \sum_{u,v \in S} d(u,v)$$

MaxMin diversification: maximize the minimum relevance (r) and dissimilarity

$$score(S) = \min_{u \in S} w(u) + \lambda \min_{u,v \in S} d(u,v)$$

Multi-criteria

Many different ways to combine

- *Maximal Marginal Relevance (MMR)* a document has high marginal relevance if it is both *relevant* to the query and contains *minimal similarity* to previously selected documents
- *Non-linear functions*: E.g., maximize the probability that an item is both relevant and diverse (e.g., non-redundant)
- Using *thresholds*

How?

Diversity: Algorithms

Most formulations of the diversity problems are NP-hard, because a *set selection problem* (set coverage)

- *Item selection at each step depends on the item selected in the previous step*

- Compute first a (relevant) result and then "diversify" it
- Produce a relevant and diverse result on the fly

Diversity: Algorithms

Interchange (swap) methods: start with the top-*k* relevant items and replace items that improve the objective function

Greedy methods: build the set incrementally, by selecting the item (or, pair of items) with the largest increase of the objective function

- Appropriate re-writing to the maxmin-maxsum dispersion problems in facility location (OR) (approximation bounds)

Diversity: Algorithms

Optimization problem

Clustering problem: *cluster items and select the centers*

Random walks on graphs

GrassHopper

Graph of items

Edge weight represents their (cosine) *similarity*

Node weight: prior *ranking* as a probability distribution *r* over the nodes (for example, based on relevance)

Parameter λ to combine the two

Random Walk with Jumps: At each step, the walker either

- with probability λ moves to a neighbor state according to similarity (the edge weights); or
- teleports to a random state according to ranking (the distribution *r)*.

One-at-a-time, the *highest rank item* is turned into an *absorbing state* and the walk is repeated

Data Diversity in Various Contexts

- Centrality measures in graphs (DivRank)
- Graph patterns
- Keyword search
- Location based queries
- Skylines queries
- …

References I (partial list) indicative

- **[AGH+09]** Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, Samuel Ieong: *Diversifying search results*. WSDM 2009: 5-14 *(example of coverage-based diversity)*
- **[GS09]** Sreenivas Gollapudi, Aneesh Sharma: *An axiomatic approach for result diversification*. WWW 2009: 381-390 *(theoretical treatment, greedy algorithms with links to the dispersion problems)*
- **[DP10]** Marina Drosou, Evaggelia Pitoura: *Search result diversification*. SIGMOD Record 39(1): 41-47 (2010) *(survey)*
- **[AK11]** Albert Angel, Nick Koudas: *Efficient diversity-aware search.* SIGMOD Conference 2011: 781-792 *(threshold-based algorithm, usefulness = probability of both relevant and diverse)*
- **[VSS+08]** Erik Vee, Utkarsh Srivastava, Jayavel Shanmugasundaram, Prashant Bhat, Sihem Amer-Yahia: *Efficient Computation of Diverse Query Results.* ICDE 2008: 228-236 *(diversity ordering of attributes, index structure)*
- **[CKC+08]** Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, Ian MacKinnon: *Novelty and diversity in information retrieval evaluation.* SIGIR 2008: 659-666 *(novelty-based diversity in IR, evaluation metrics)*
- **[CCS+11]** Charles L. A. Clarke, Nick Craswell, Ian Soboroff, Azin Ashkan: *A comparative analysis of cascade measures for novelty and diversity.* WSDM 2011: 75-84 *(IR diversity-aware metrics)*
- **[CG98]** Jaime G. Carbonell, Jade Goldstein: *The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries*. SIGIR 1998: 335-336 *(seminal paper on MMR)*

References II (partial list)

- **[ZMK+05]** Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen: *Improving recommendation lists through topic diversification.* WWW 2005: 22-32 *(assumes taxonomy of topics, evaluation)*
- **[VC11]** Saul Vargas, Pablo Castells: *Rank and relevance in novelty and diversity metrics for recommender systems.* RecSys 2011: 109-116 *(various aspects of diversity and metrics, discovery-choice-relevance aspects)*
- **[YLA09]** Cong Yu, Laks V. S. Lakshmanan, Sihem Amer-Yahia: *It takes variety to make a world: diversification in recommender systems.* EDBT 2009: 368-378 *(diversification based on dissimilarity of explanations associated with each recommended item)*
- **[BLY12]** Allan Borodin, Hyun Chul Lee, Yuli Ye: *Max-Sum diversification, monotone submodular functions and dynamic updates.* PODS 2012: 155-166 *(approximation bounds for the maxsum problem using submodularity)*
- **[CZS+12]** Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, Tamas Jambor: *Auralist: introducing serendipity into music recommendation.* WSDM 2012: 13-22 *(serendipity, nice treatment of various aspects of diversity)*
- **[ZGG+07]** Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, David Andrzejewski: *Improving Diversity in Ranking using Absorbing Random Walks.* HLT-NAACL 2007: 97-104 *(the GrassHopper algorithm)*
- **[VRB+11]** Marcos R. Vieira, Humberto Luiz Razente, Maria Camila Nardini Barioni, Marios Hadjieleftheriou, Divesh Srivastava, Caetano Traina Jr., Vassilis J. Tsotras: *On query result diversification.* ICDE 2011: 1163-1174 *(comparison of various algorithms, proposal of "randomized" greedy)*
- **[TTH+15]** Duong Chi Thang, Nguyen Thanh Tam, Nguyen Quoc Viet Hung, Karl Aberer: *An Evaluation of Diversification Techniques.* DEXA (2) 2015: 215-231 *(experimental evaluation of algorithms)*

Our work

r-DisC set: r-Dissimilar and Covering set

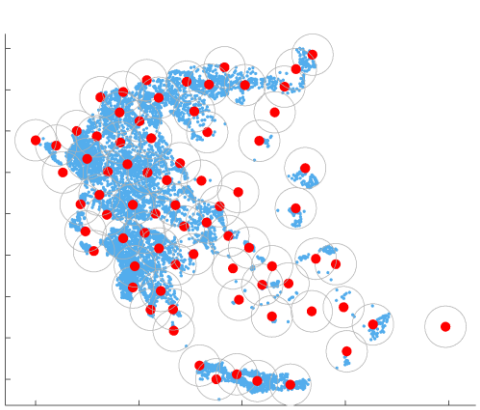
What is the right size for the diverse subset *S*? What is a good *k*?

What if… instead of *k*, a *radius r*

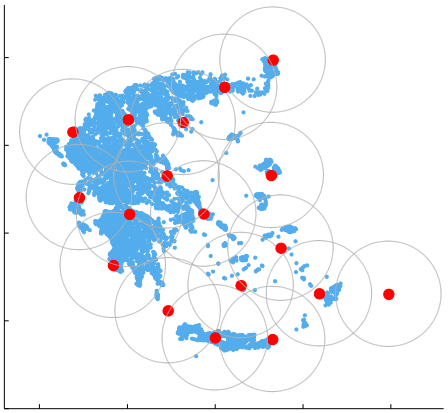
Select a representative subset $S \subseteq P$ such that:

1. *For each item p in P, there is at least one similar item p' in S, d(p, p') <= r* ***(coverage)***
2. *No two items p, p' in the diverse set S are similar with each other, d(p, p') > r* ***(dissimilarity)***

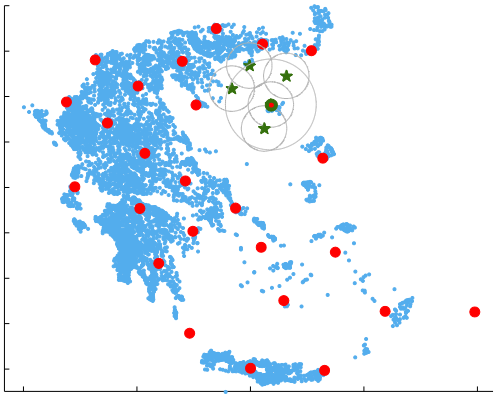
r-DisC set: r-Dissimilar and Covering set



Zoom-in



Zoom-out



Local zoom

- Small *r*: more and less dissimilar items (*zoom in*)
- Large *r*: less and more dissimilar items (*zoom out*)
- Local zooming at specific items

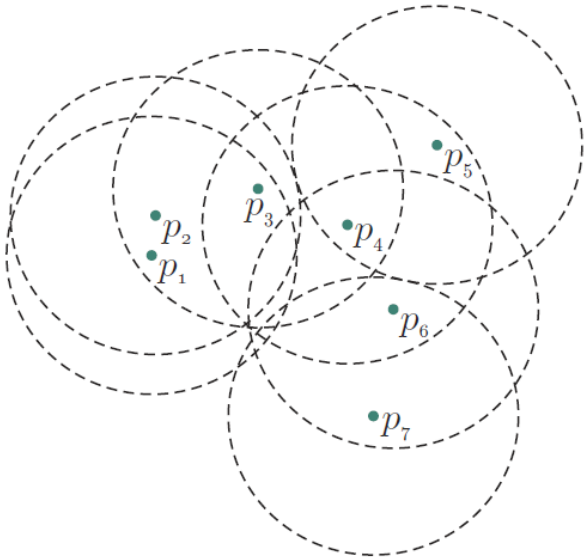
r < smallest distance, |S| = n

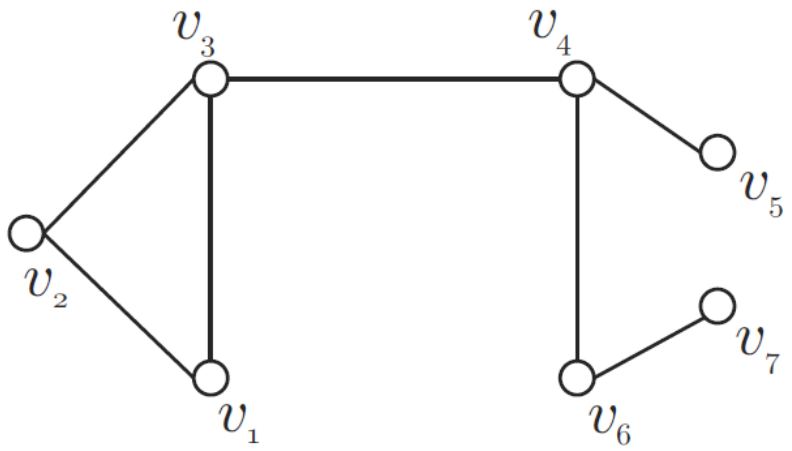
r > largest distance, |S| = 1

Graph Model

Model the problem as a *graph*

- Items are nodes
- There is an edge between two nodes, if distance ≤ r



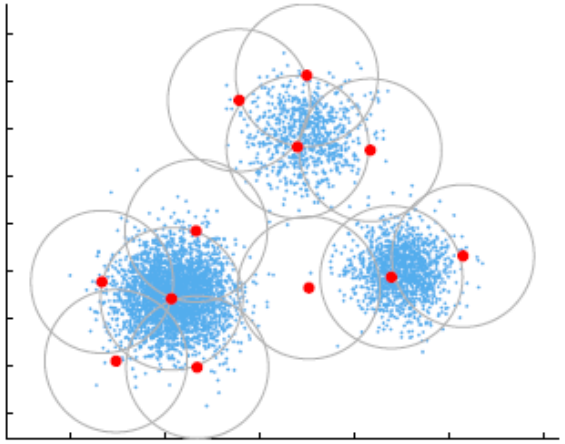


Equivalent to finding a minimal

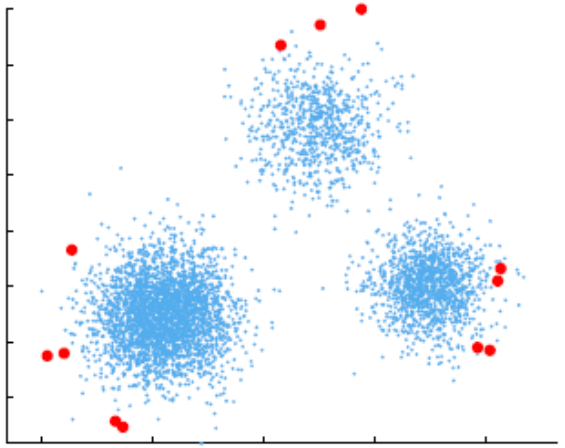
- Independent (no edge about nodes in the set) and
- Dominating (all nodes outside connected with at least one inside)

subset of the corresponding graph (aka maximal independent subset)

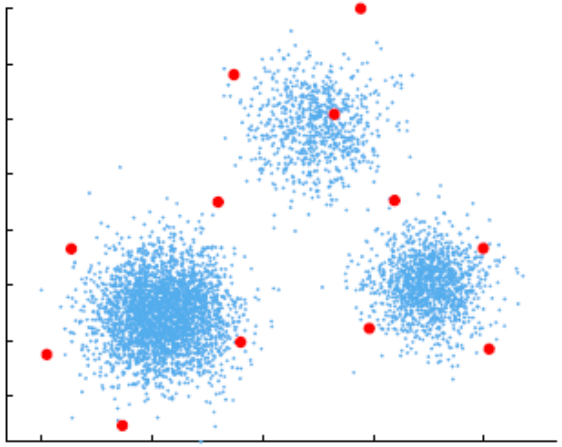
Comparison with other models



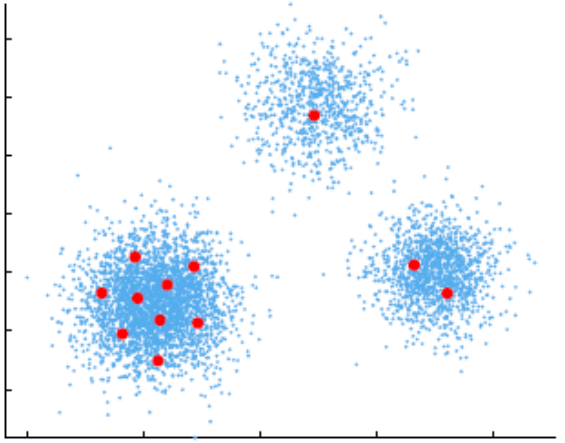
r-DisC



MAXSUM



MAXMIN



k-medoids

Zooming

User ***interactively*** change the radius *r* to *r'* and compute a new diverse set

- *r'* < *r*: zoom-in
- *r'* > *r*: zoom-out

Two requirements:

1. Support an incremental mode of operation:
   - the new set should be as close as possible to the already seen result
2. The size of the new set should be as close as possible to the size of the minimum r'-DisC diverse subset

There is no subset relation between the r-DisC diverse and the r'-DisC diverse subsets of a set of objects P (the two sets may be completely different)

DisC-Extensions

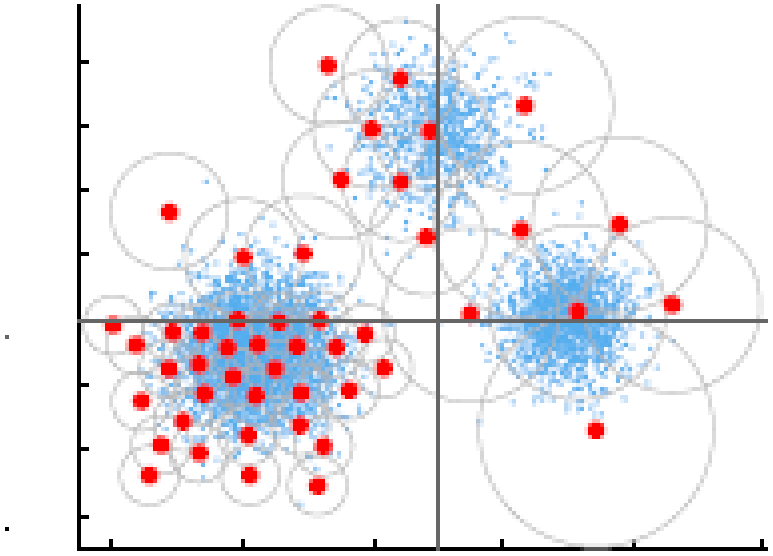
Different radii per item

Radius as a function of the item

- ✓ Based on importance
- ✓ Based on relevance

Directed graph

- In general, there may be no solution
- In our case, constructive proof there exists



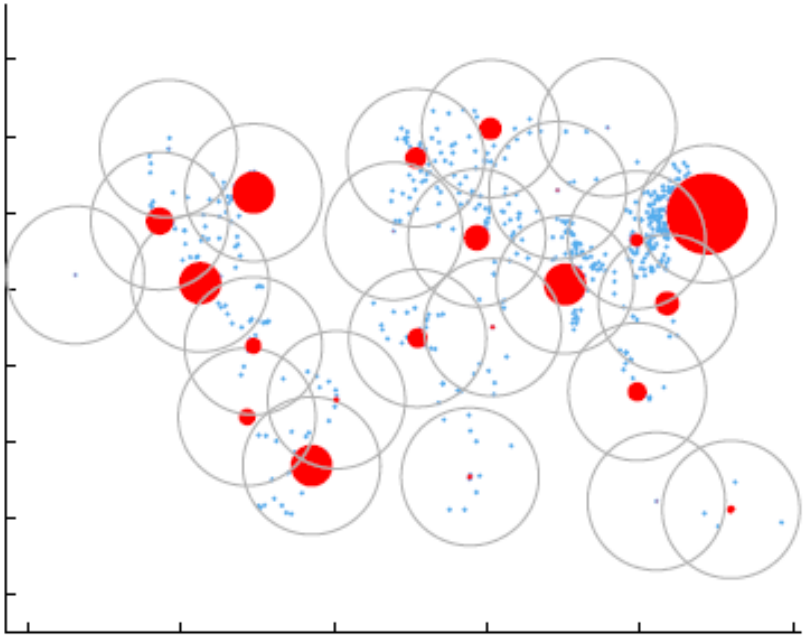
DisC-Extensions

Different weight per point

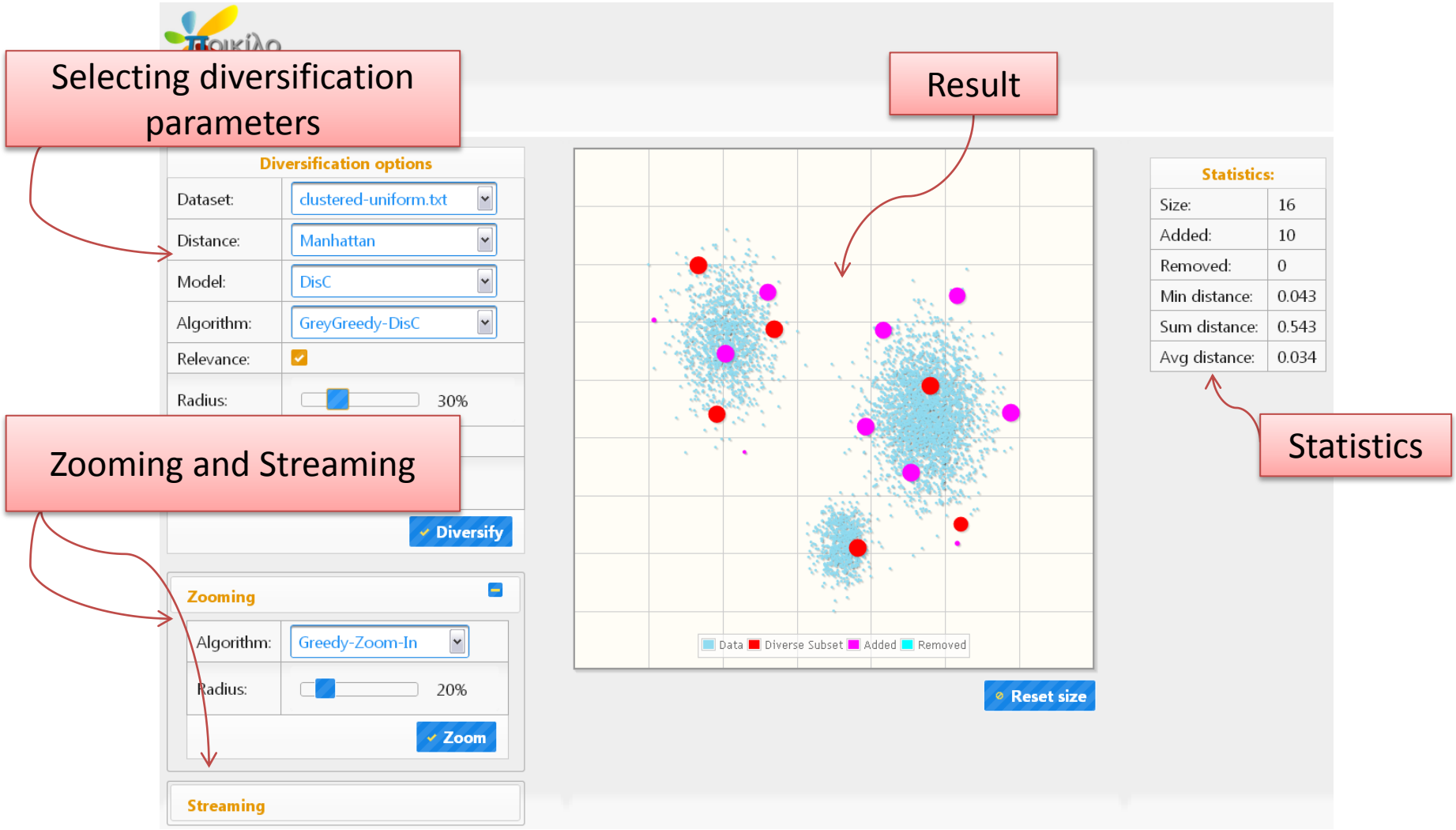
Find the r-DiSC set with the minimum

$$f(S) = \sum_{p_i \in S} \frac{1}{w(p_i)}$$

When all weights are equal, the problem is reduced to finding a minimum r-DisC subset



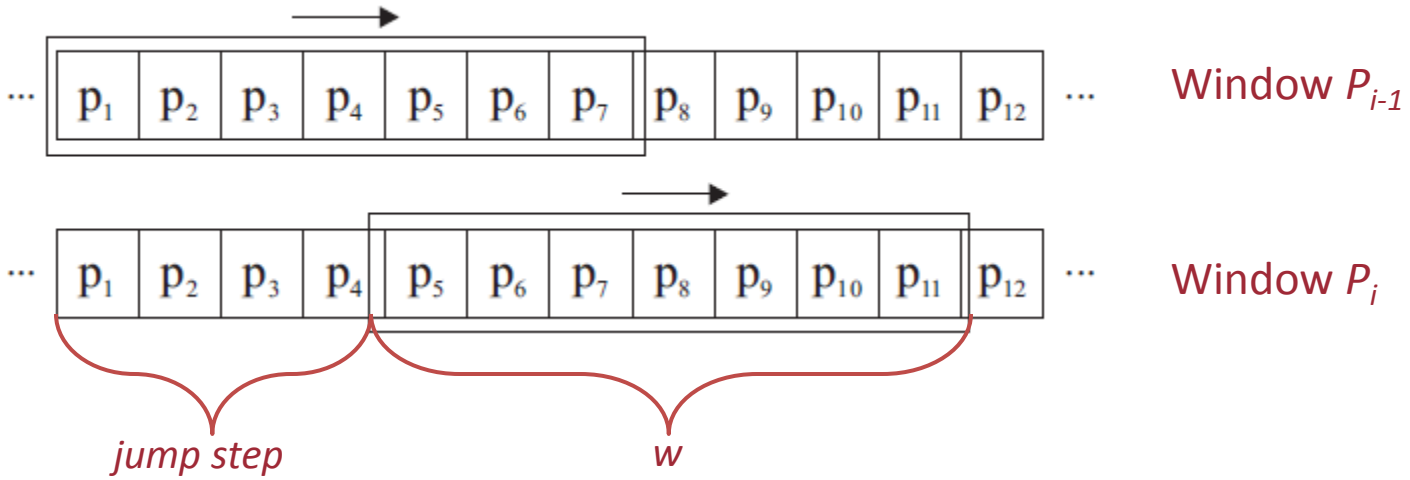
Visualizing Diverse Items



Diversity over Dynamic Sets

We study the dynamic/streaming diversification problem:

- New items (books, movies etc.) are added to a recommender system.
- News apartments become available while old ones are not available any more.
- Microblogging applications (e.g., twitter)



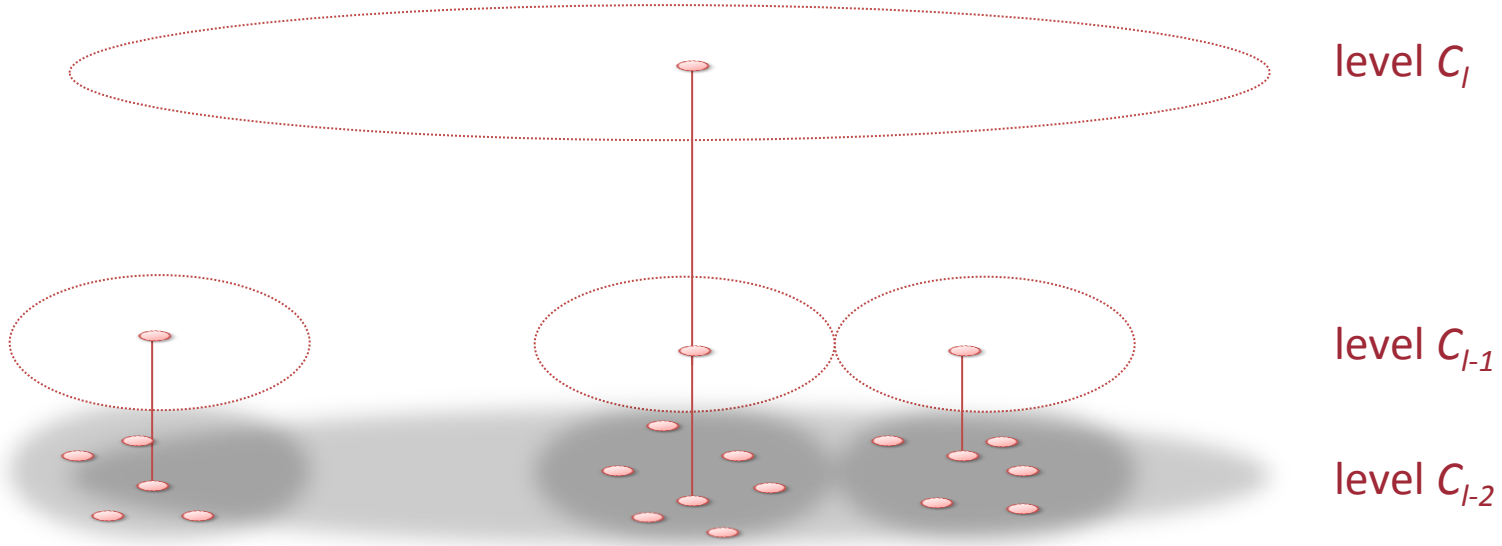
- New items arrive and older items expire (window jumps, e.g., consequent logins)
- We want to provide users with a continuously updated subset of the top-*k* most diverse recent items in the stream.

Indexing

We index items in *P* using a cover tree*

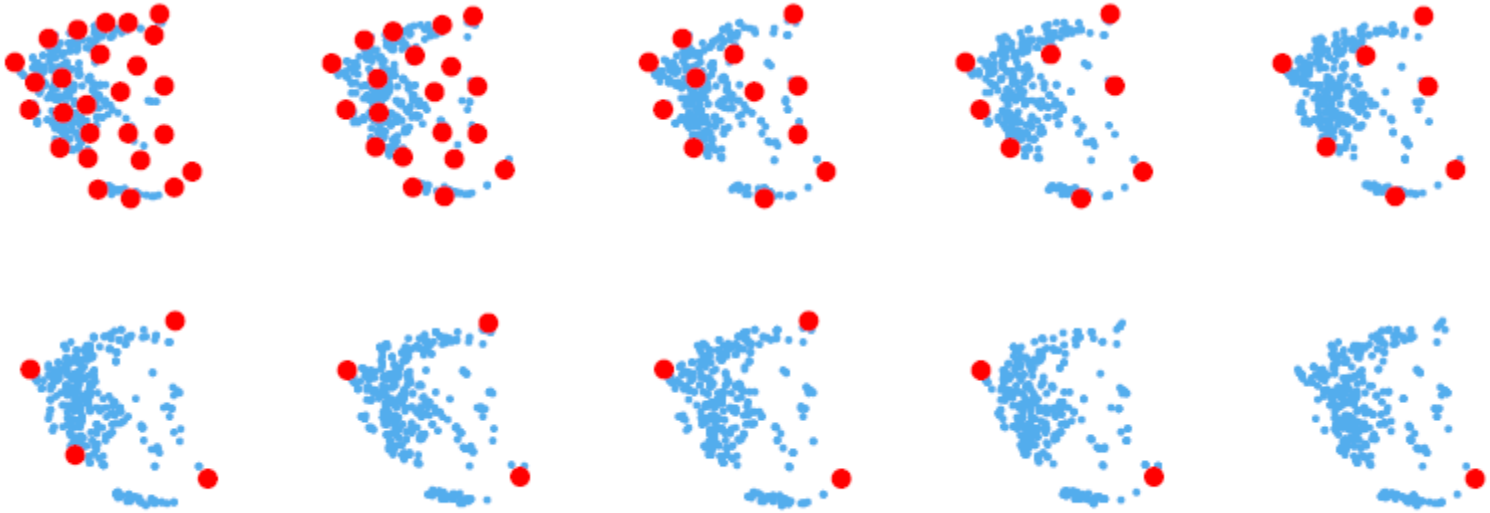
Cover tree:

- Leveled tree: Lowest level <- items in *P*
- Levels are numbered, e.g., -4 (leaf), -3, …, 0, … 3, .. 5 (root) and each level is a "cover" for all levels beneath it
- Items at higher levels are farther apart from each other than items at lower levels.



* [BKL06] A. Beygelzimer, S. Kakade, and J. Langford. *Cover Trees for Nearest Neighbor*. ICML, 2006.

Cover Tree: Example of some levels

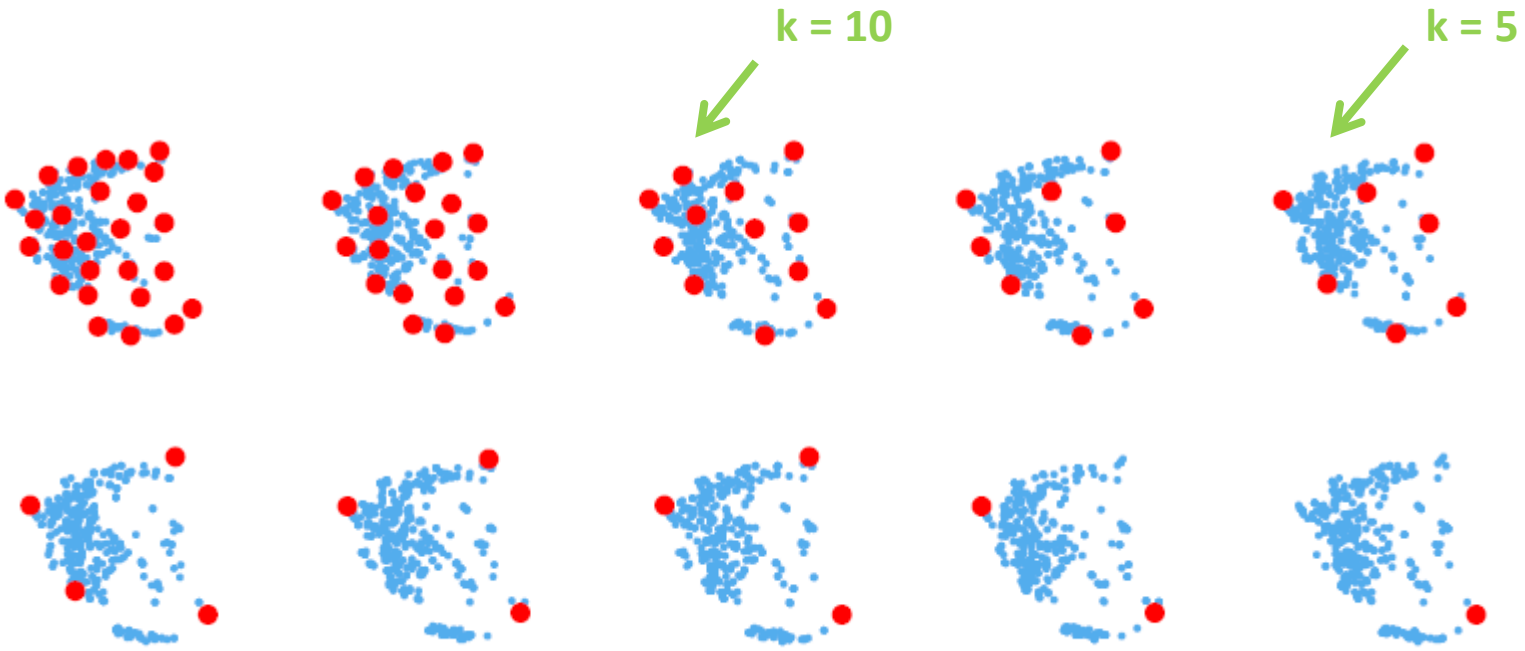


Example: higher levels of a cover tree for cities in Greece, where distance is their geographical distance

Cover Tree: Diversity computation

The Level Family of Algorithms

Basic Idea: Select *k* distinct items from the highest possible level



Scalability: depend on the size of the level not on the size of the dataset

DisC Diversity

Marina Drosou, Evaggelia Pitoura: *Multiple Radii DisC Diversity: Result Diversification Based on Dissimilarity and Coverage.* ACM Trans. Database Syst. 40(1): 4 (2015)

Marina Drosou, Evaggelia Pitoura: *DisC diversity: result diversification based on dissimilarity and coverage.* PVLDB 6(1): 13-24 (2013) *(Best paper award)*

Diversity in Streams

Marina Drosou, Evaggelia Pitoura: *Diverse Set Selection Over Dynamic Data.* IEEE Trans. Knowl. Data Eng. 26(5): 1102-1116 (2014)

Marina Drosou, Evaggelia Pitoura: *Dynamic diversification of continuous data.* EDBT 2012: 216-227

Marina Drosou, Kostas Stefanidis, Evaggelia Pitoura: *Preference-aware publish/subscribe delivery with diversity.* DEBS 2009

Summary

- Diversity (coverage, dissimilarity, novelty, serendipity) *improves the value* of data
- *DisC diversity* provides a zoom-able view of a data set that ensures both coverage and dissimilarity
- Diversity of *streaming data* adds the dimension of time

Diversity in Social Networks

Homophily

"Όμοιος ομοίω αεί πελάζει" (Plato)

"Birds of a feather flock together"

Caused by two related social forces

- *Selection:* People seek out similar people to interact with
- *Social influence:* *People become similar to those they interact with*

Both processes contribute to homophily and lack of diversity, but

- Social influence leads to community-wide homogeneity
- Selection leads to fragmentation of the community

Opinion Formation

Complex process: many models

Commonly-used opinion-formation model (of Friedkin and Johnsen, 1990) (opinion – real number)

- Each individual *i* has an innate and an expressed opinion.
- At each step updates her expressed opinion
  - adheres to her innate opinion with a certain weight a_i and
  - is *socially influenced* by its neighbors with a weight $1-a_i$

Opinion Formation

An opinion formation process is polarizing if it results in increased divergence of opinions.

Empirical studies have shown that homophily results in polarization.

A past Λ14 project

Diversify opinions within communities

Select a set of *k* individuals to influence so that they "change" opinions

Create a set of *k* new connections between nodes in different communities with contrasting views

Debiasing the Wisdom of the Crowd

- Wisdom of the crowd (collective wisdom): aggregation of information in groups, results in decisions often better than by any single member of the group.

- When individuals become aware of the estimates of others, they may revise their own estimates

  Experimental evidence that this holds *also for factual questions* and monetary incentives: Groups were initially "wise," knowledge about estimates of others narrows the diversity of opinions

Debiasing the Wisdom of the Crowd

- Take into account *the effect of social influence* when estimating the collective wisdom of a crowd
  - Efficient sampling for innate opinions
  - Since only the expressed opinion of the nodes (cannot directly observe their innate opinion), algorithms need to take care of debiasing the expressed opinions of the nodes that they sample.

J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci.* USA, 108(22), 1990
Abhimanyu Das, Sreenivas Gollapudi, Rina Panigrahy, Mahyar Salek: *Debiasing social wisdom*. KDD 2013

Opinion Diversity in Crowdsourcing Markets

Similarity-driven Model (S-Model)

No specific query/task

Given the similarity of workers maximize their average diversity (MAXAVG)

Task-driven model (T-Model)

Specific query/task

- Model the opinion of each worker as a probability ranging from 0 to 1 (indicating opinions from negative to positive)
- A user specifies a required number of workers with positive and negative opinions.
- Maximize the probability that the user's demand is satisfied.

Ting Wu, Lei Chen, Pan Hui, Chen Jason Zhang, Weikai Li: Hear the Whole Story: Towards the Diversity of Opinion in Crowdsourcing Markets. PVLDB 8(5): 485-496 (2015)

Diversity, Fairness, Responsibility

Diversity of data and opinions

How does diversity of data presented to individuals or groups affects the fairness of their decision?

Lack of (opinion, data) diversity leads to polarization and bias?

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. *Exposure to Ideologically Diverse News and Opinion on Facebook.* Science 348:1130–1132, 2014

Stages in Facebook Exposure Process

1. *Friends network*: ideological homophily
2. *News feed*: more or less diverse content with algorithmically ranked News Feed
3. *Users' choices*: click through to ideologically discordant content.

News Feed Ranking

"The *order* in which users see stories in the News Feed depends on *many factors*, including
how often the viewer visits Facebook,
how much they interact with certain friends, and
how often users have clicked on links to certain websites in News Feed in the past."

Dataset: users

10.1 million *active* U.S. users *who self-report* their ideological affiliation

All Facebook users can self-report their political affiliation, **9%** of U.S. over 18

Dataset: content

7 million distinct Web links (URLs) shared by U.S. users over a 6-month period between 7 July 2014 and 7 January 2015

Classified stories as

- **Hard content** (such as national news, politics, or world affairs) or
- **Soft content** (such as sports, entertainment, or travel)

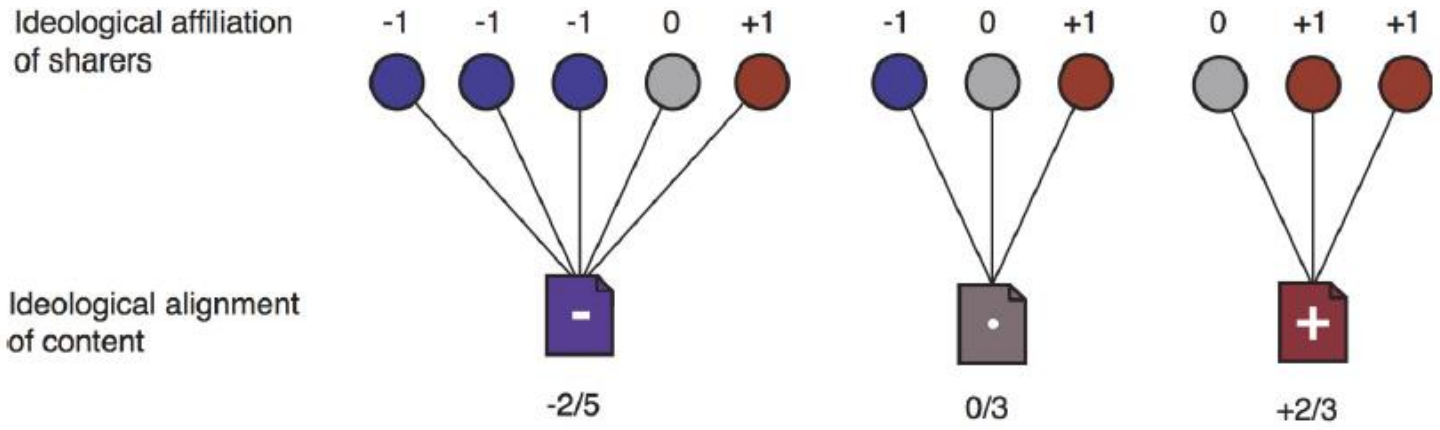
by training a *support vector machine* on unigram, bigram, and trigram text features

Approximately **13%** hard content.

226,000 distinct hard-content URLs shared **by at least 20 users** who volunteered their **ideological affiliation** in their profile

Labeling stories (content alignment)

measure *content alignment (A)* for each hard story:
average of the ideological affiliation of each user who shared the article.

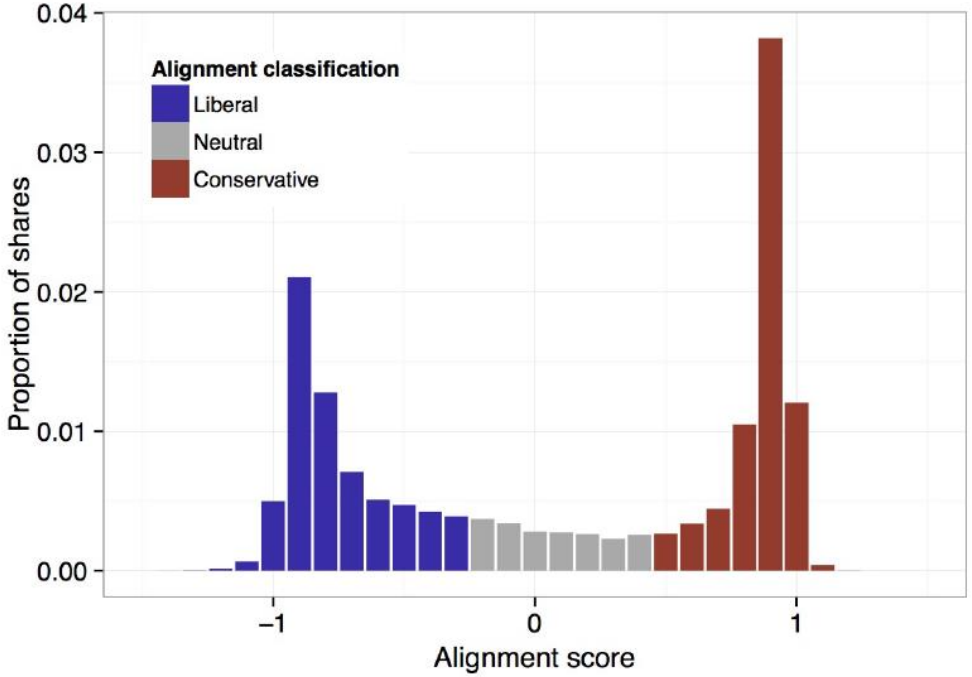


- measure of the *ideological alignment of the audience* who shares an article, *not a measure of political bias or slant of the article*

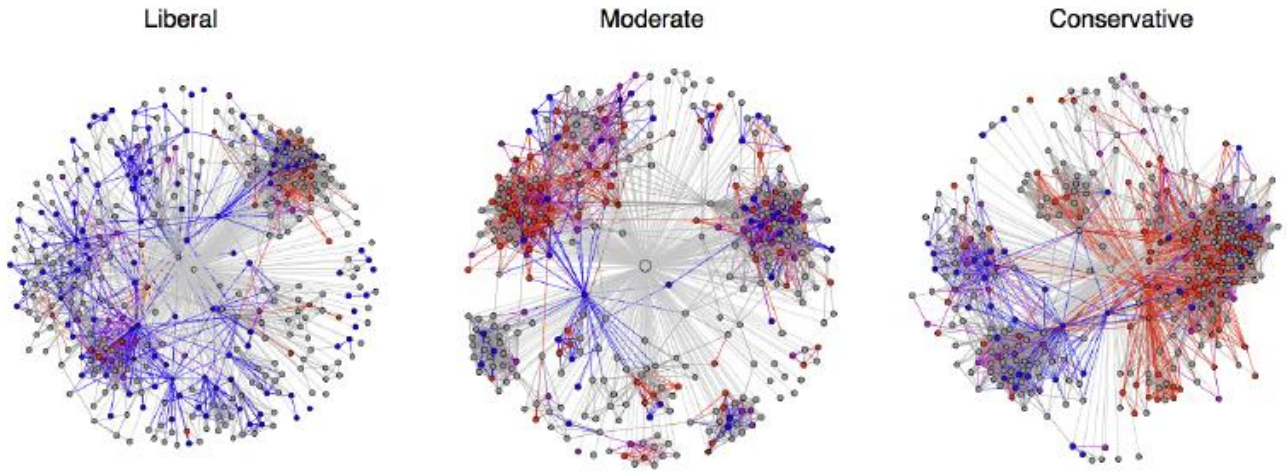
Labeling stories (content alignment)

FoxNews.com is aligned with conservatives (As = +.80)
HuffingtonPost.com is aligned with liberals (As = -.65)

Substantial polarization

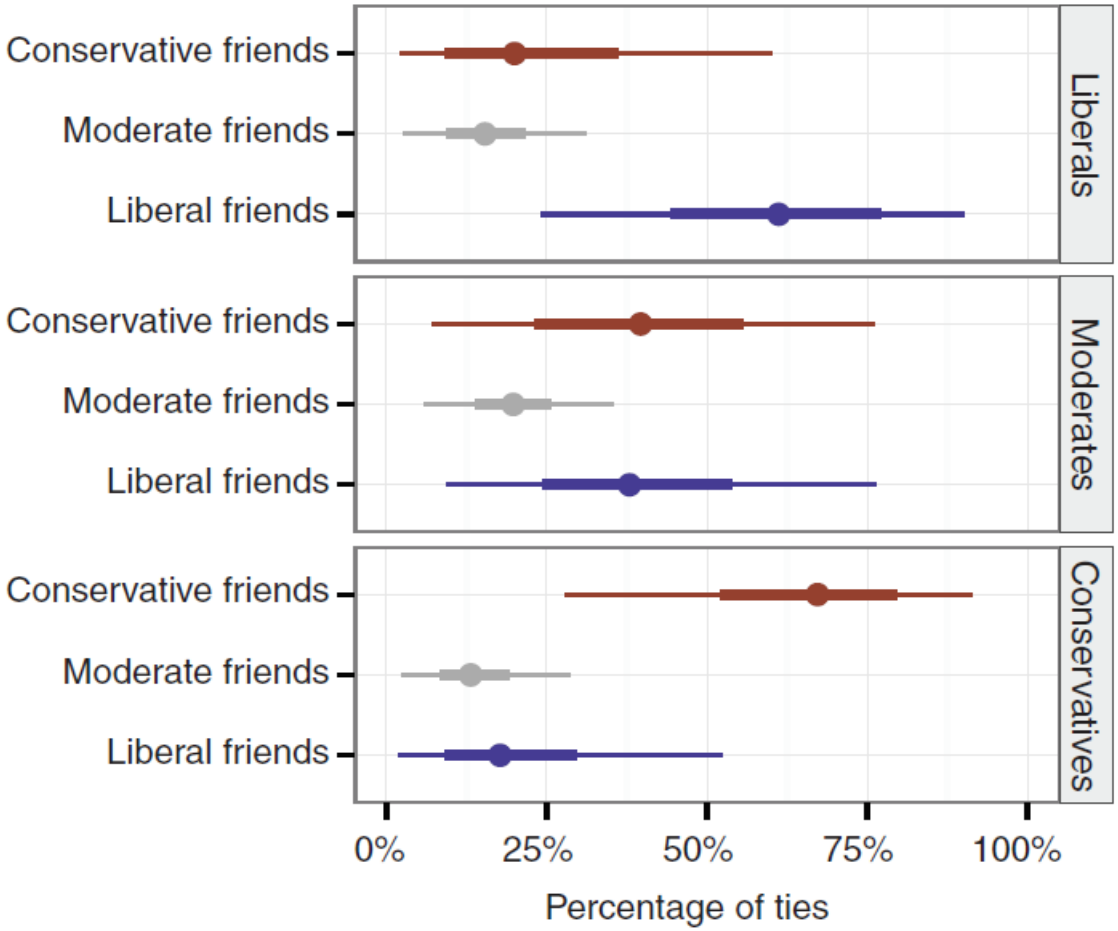


Homophily in the Friends Network



Example social networks for a liberal, a moderate, and a conservative. Points are individuals' friends, and lines designate friendships between them.

Homophily in the Friends Network



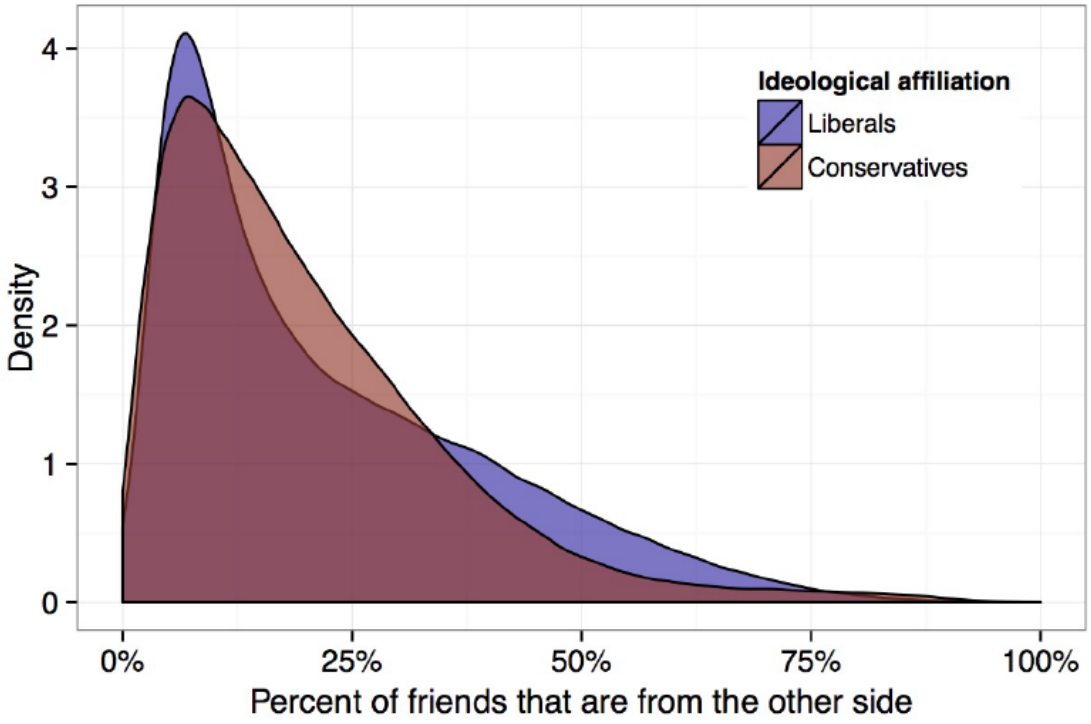
Median proportion of friendships

- of liberals with conservatives 0.20,
- of conservatives maintain with liberals 0.18

Homophily in the Friends Network

On average, about 23 percent of their friends report an affiliation on the opposite side
A wide range of network diversity

- 50% between 9 and 33 percent,
- 25% less than 9 percent
- 25% more than 33 percent



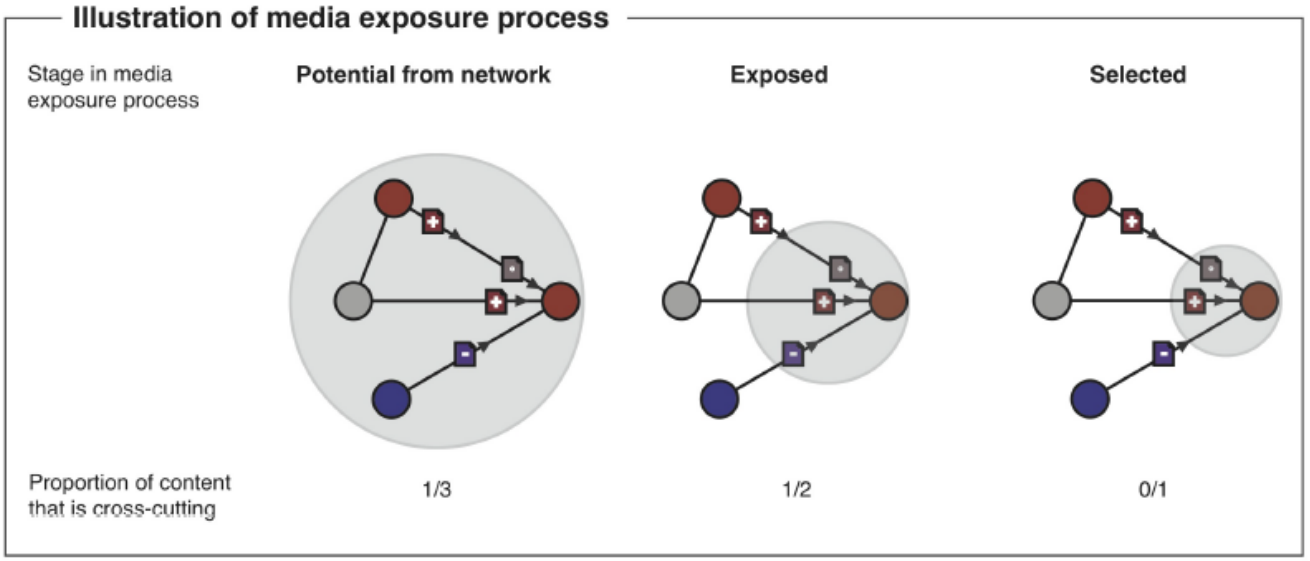


Illustration of how the exposure process consists of three phases: (1) the news your friends share (Potential from network), (2) ranking and the time that individuals take to scroll governs what they see in their News Feeds (Exposed), (3) clicking through to actual article (Selected).

Content shared by friends

If from **random** others,
~45% cross-cutting for liberals
~40% for conservatives

If from **friends**,
~24% crosscutting for liberals
~35% crosscutting for conservatives

News Feed

After ranking, there is on average *slightly less crosscutting*

risk ratio of *x* percent:
people were *x* percent less likely to see crosscutting articles that have been shared by friends, compared to the likelihood of seeing ideologically consistent articles that have been shared by friends.

risk ratio

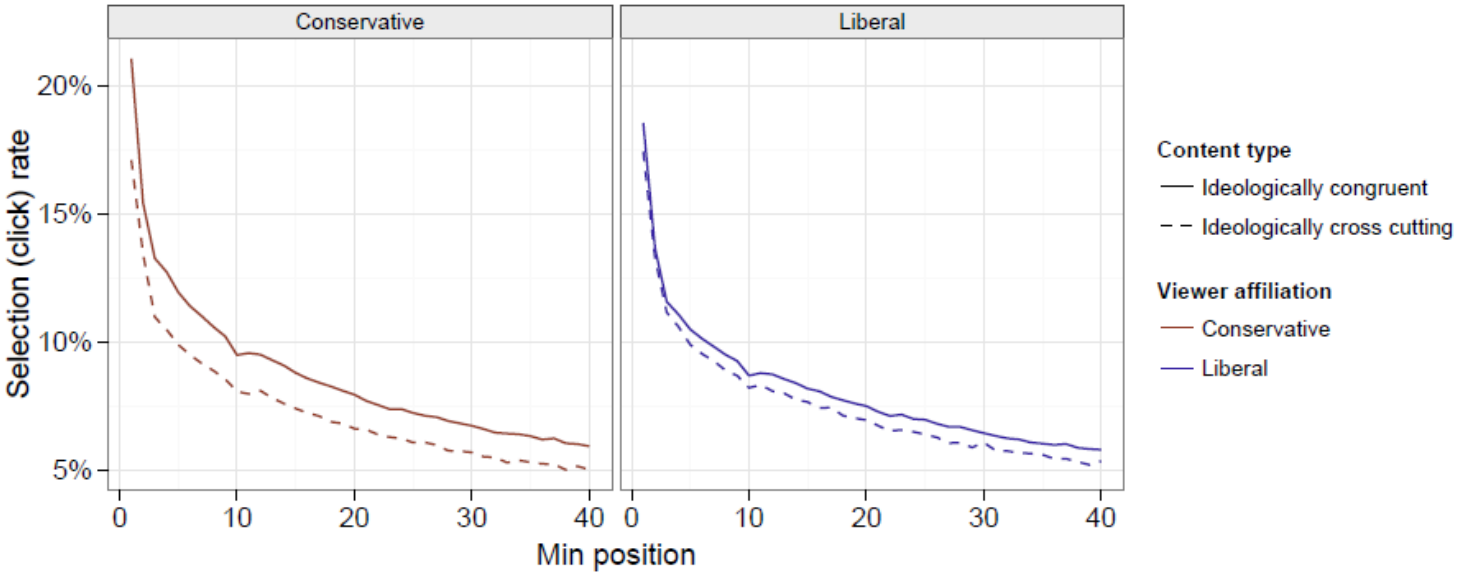
- 5% for conservatives
- 8% for liberals

Clicked

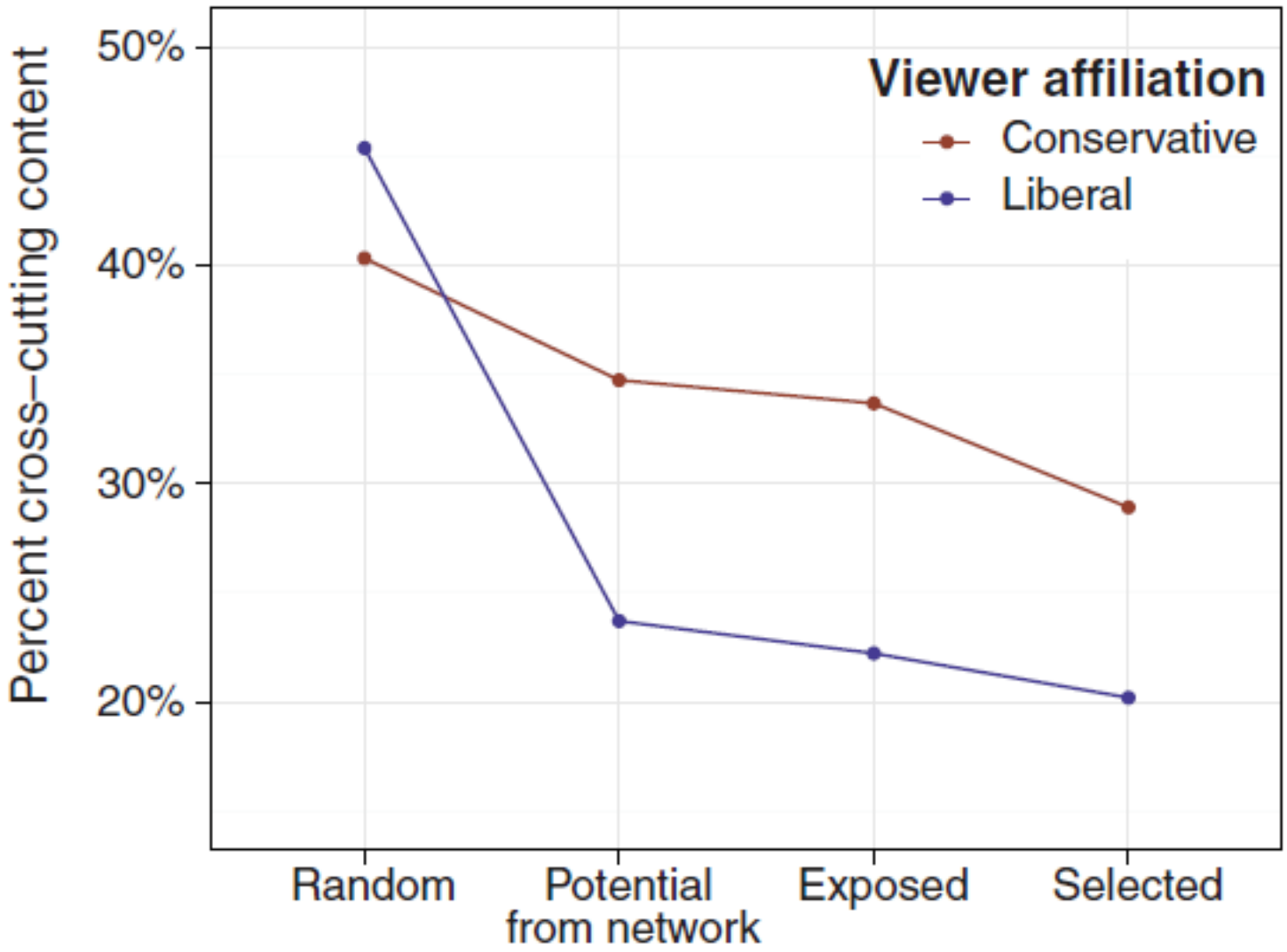
Risk ratio
17% for conservatives
6% for liberals,

On average, viewers clicked on 7% of hard content available in their feeds

Clicked



the click rate on a link is negatively correlated with its position in the News Feed


Percent cross–cutting content
50%
40%
30%
20%
Viewer affiliation
Conservative
Liberal
Random
Potential from network
Exposed
Selected

Limitation (as described by the authors)

- Limited to active users who volunteer an ideological affiliation
- Facebook users tend to be younger, more educated, and more often female as compared with the U.S. population as a whole
- Other forms of social media, such as blogs or Twitter, different patterns of homophily among politically interested users (largely because ties tend primarily to form based on common topical interests and/or specific content, whereas Facebook ties primarily reflect many different offline social contexts: school, family, social activities, and work, which favor cross-cutting social ties
- Distinction between exposure and consumption is imperfect; individuals may read the summaries of articles that appear in the News Feed and therefore be exposed to some of the articles' content without clicking through.

A WSJ site

Blue Feed, Red Feed site
See Liberal Facebook and Conservative Facebook, Side by Side

Based on the reactions by conservative/liberals as in the paper

http://graphics.wsj.com/blue-feed-red-feed/

Questions?