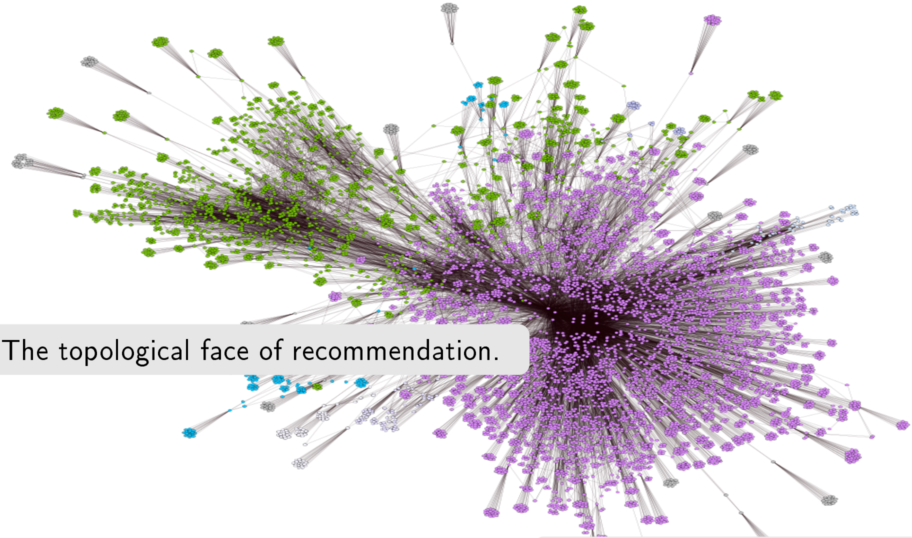
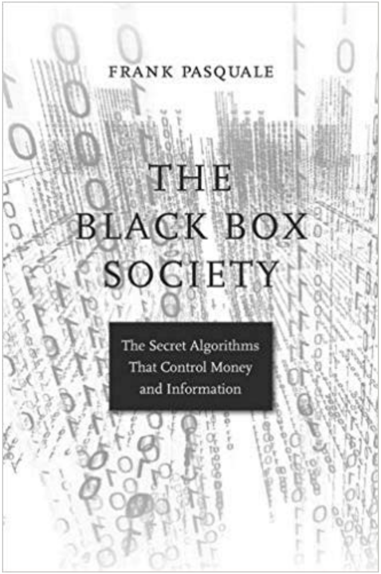


The topological face of recommendation.

Erwan Le Merrer, *Inria*
Gilles Trédan, *LAAS/CNRS*

The "black box-ization" of interactions



The “black box-ization” of interactions

VERS L’AUTOMATISATION DE LA CENSURE POLITIQUE - FÉLIX TRÉGUER

« L’urgence, c’est de rompre l’alliance des appareils policiers et des grands marchands d’infrastructures numériques »

paru dans lundimatin#180, le 26 février 2019

Nous publions ici un article généreusement transmis par nos confrères de La Quadrature du Net sur les nouvelles formes de censure politique dans l’espace virtuel : grâce à l’intelligence artificielle, des milliers de contenus soi-disant « terroristes » postés sur facebook ou youtube sont automatiquement supprimés chaque jour. Pour cela, les États, loin d’être concurrencés par les géants de l’internet, collaborent bien plutôt avec eux, notamment en légiférant pour aménager la possibilité d’une censure extra-judiciaire (suppression automatique des contenus).

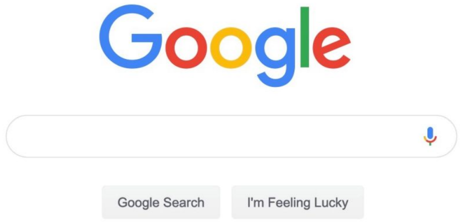
Nous sommes à un tournant de la longue histoire de la censure. Ce tournant, c’est celui de la censure privée et automatisée. Il acte une rupture radicale avec les garanties associées à la liberté d’expression que les luttes démocratiques du XIXe siècle nous avaient léguées en héritage.

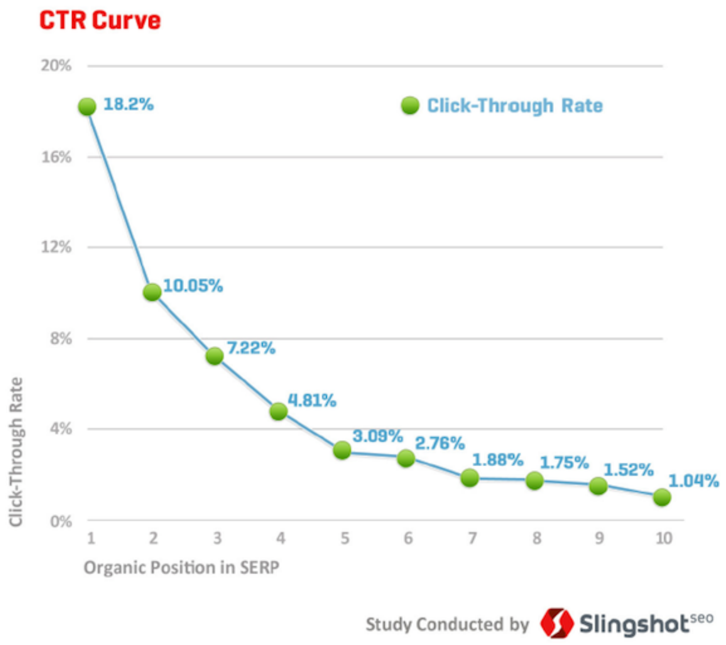
APPEL À DONS

Computers good old days

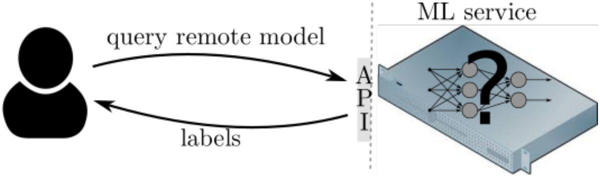
Internet good old days

Today: “oracle”-like services



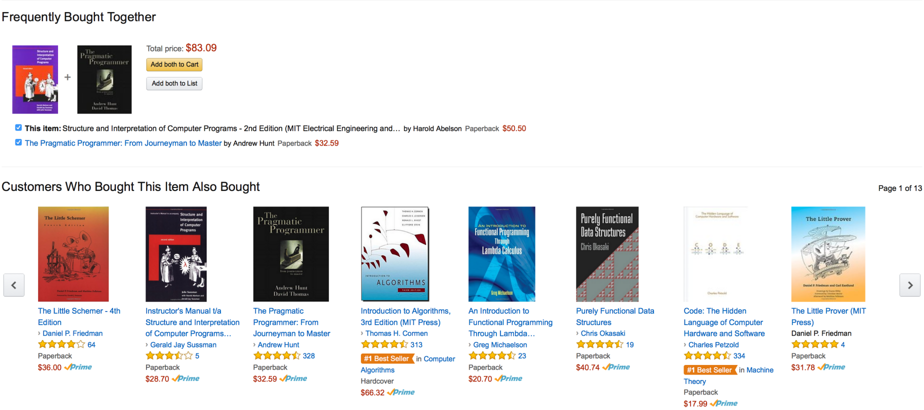

CTR Curve
20%
16%
12%
8%
4%
0%
Click-Through Rate
18.2%
10.05%
7.22%
4.81%
3.09%
2.76%
1.88%
1.75%
1.52%
1.04%
Click-Through Rate
1
2
3
4
5
6
7
8
9
10
Organic Position in SERP
Study Conducted by Slingshot seo

Turning point to the black box era



- Input: user actions/data. Arbitrary processing: output/results
- Users cannot access the data, history, algorithm...
- Trust given to the remote service/algorithm,
  - while it has big interest in manipulating the outputs (e.g., ads)

Example 2: Recommendations (Gilles's talk)



Example 2: Recommendations (Gilles's talk)

Amazon is huge. The ecommerce giant accounted for 43% of 2016 online retail sales in the US, according to Slice Intelligence. With its latest acquisition of Whole Foods and its foray into cashless shopping with Amazon Go, Amazon looks set to assert its dominance in the physical retail space as well.

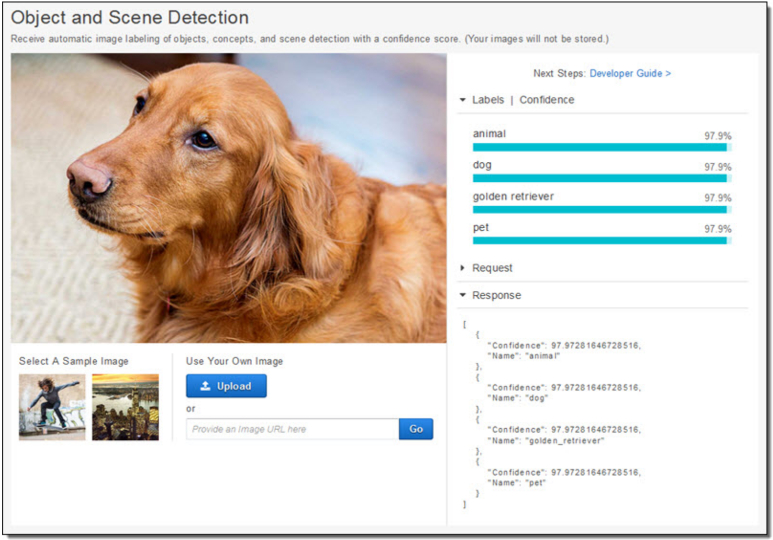
Many factors contribute to Amazon's success, but recently, artificial intelligence (AI) is increasingly being touted as a key pillar of Amazon's competitive advantage. And one of Amazon's best applications of AI is in its on-site product recommendations.

Amazon strives to create a personalized shopping experience for every customer. In a page titled 'Your Amazon.com', users are recommended a unique selection of products based on their past shopping behavior. According to research by McKinsey, a mind-boggling 35% of Amazon's sales come from such recommendations.

Example 3: Credit scoring

- Nowadays: default prediction by models $\rightarrow$ score $\rightarrow$ decision
- Data: thousands of factors, do you know/understand them all?

Example 4: From image classification APIs ...



Example 4: ... to self driving cars

(a) Input 1 (b) Input 2 (darker version of 1)

Figure 1: An example erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a). [1]

1. DeepXplore @ SOSP 2017

FOOLING THE AI

Deep neural networks (DNNs) are brilliant at image recognition — but they can be easily hacked.

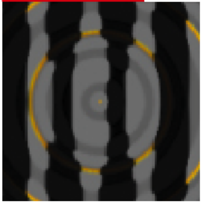
These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.

Stop

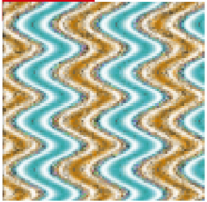
Speed limit 45

Scientists have evolved images that look like abstract patterns — but which DNNs see as familiar objects.

King penguin



Starfish



… to the infamous social credit

Our near future, the cybernetic dream?

Current solutions fail

- **Explainability**: good only if you access the algorithm **locally**!

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$) [2]

2. LIME: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016
3. https://www.gouvernement.fr/argumentaire/le-gouvernement-publie-le-code-des-algorithmes-de-parcoursup

Current solutions fail

- **Explainability**: good only if you access the algorithm **locally**!

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$) [2]

- **Transparency**: "please trust me I am clean"

Le Gouvernement publie le code des algorithmes de Parcoursup

Une première à l'échelle de l'État : le Gouvernement a publié le 21 mai 2018, le code informatique du cœur algorithmique de la plateforme d'orientation universitaire Parcoursup. [3]

2. LIME: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016
3. https://www.gouvernement.fr/argumentaire/le-gouvernement-publie-le-code-des-algorithmes-de-parcoursup

https://framagit.org/parcoursup/algorithmes-de-parcoursup/-/blob/master/java/parcoursup/propositions/algo/Voeu.java

Projects Groups Snippets Help

algorithmes-de-parcoursup

- Project overview
- **Repository**
  - **Files**
  - Commits
  - Branches
  - Tags
  - Contributors
  - Graph
  - Compare
- Issues 5
- Merge Requests 2
- Analytics
- Wiki
- Snippets
- Members

Parcoursup > algorithmes-de-parcoursup > **Repository**

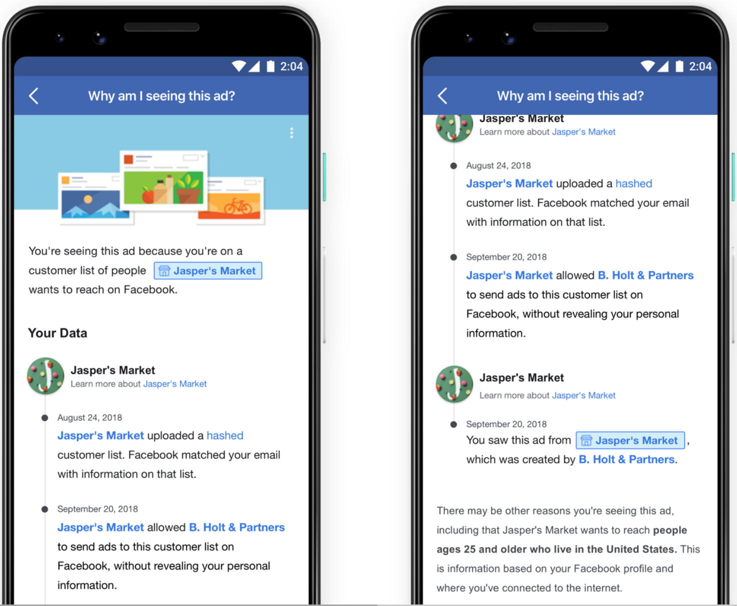
master | algorithmes-de-parcoursup / java / parcoursup / propositions / algo / **Voeu.java**

évolutions 2019: répondeur automatique, meilleurs bacheliers, affichages candidats
Parcoursup authored 8 months ago

Voeu.java 11.1 KB

```java

/* Copyright 2018 © Ministère de l'Enseignement Supérieur, de la Recherche et de
l'Innovation,
    Hugo Gimbert (hugo.gimbert@enseignementsup.gouv.fr)

    This file is part of Algorithmes-de-parcoursup.

    Algorithmes-de-parcoursup is free software: you can redistribute it and/or modify
    it under the terms of the Affero GNU General Public License as published by
    the Free Software Foundation, either version 3 of the License, or
    (at your option) any later version.

    Algorithmes-de-parcoursup is distributed in the hope that it will be useful,
    but WITHOUT ANY WARRANTY; without even the implied warranty of
    MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
    Affero GNU General Public License for more details.

    You should have received a copy of the Affero GNU General Public License
    along with Algorithmes-de-parcoursup.  If not, see <http://www.gnu.org/licenses/>.

 */
package parcoursup.propositions.algo;

import javax.xml.bind.annotation.XmlTransient;
import parcoursup.propositions.meilleursbacheliers.MeilleurBachelier;

public class Voeu {

    /* caractéristiques identifiant de manière unique le voeu dans la base de données */
    public final VoeuUID id;

    /* groupe d'affectation du voeu */
    @XmlTransient
```

Potentially adversarial algorithms: beware of "fair-washing"



Potentially adversarial algorithms: beware of "fair-washing"

Our experiments demonstrated that Facebook's ad explanations are often incomplete and sometimes misleading, and that Facebook's data explanations are incomplete and often vague. These findings have important implications for users, as they may lead them to incorrectly conclude how they were targeted with ads. Moreover, these findings also suggest that malicious advertisers may be able to obfuscate their true targeting attributes by hiding rare (and potentially sensitive) attributes by also selecting very common ones. To make matters worse, Twitter recently introduced explanations that are similar to Facebook's explanations. This underscores the urgent need to provide properly designed explanations as social media advertising services mature. We hope that our study will provide a basis to guide such a design. [4]

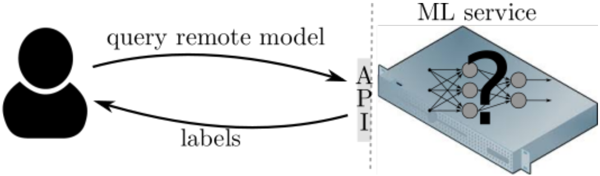
4. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations, NDSS 2018

Potentially adversarial algorithms: beware of "fair-washing"

The bouncer problem![5]

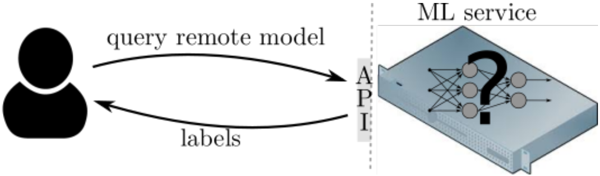
5. The bouncer problem: challenges for remote explainability, arXiv 2019

Researchers, hackers: we need audit algorithms



- General framework for user-sided audits:
  - tweak craftable input
  - submit to the black-box
  - collect results
  - if enough to conclude on hypothesis: return
  - loop;

Researchers, hackers: we need audit algorithms



- General framework for user-sided audits:
  - tweak craftable input
  - submit to the black-box
  - collect results
  - if enough to conclude on hypothesis: return
  - loop;

BUT assuming that the black-box can be **adversarial**

AND that the number of submissions **must be small**

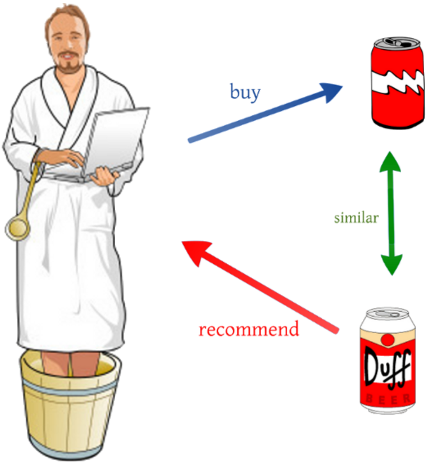
The black-box society looks quite real

- From **user-control of algorithms to algorithmic-control**
- Huge impact, close to **no tools today** to assess this
- We need user-sided **audit algorithms**
  - Blend of security, data science, behavioural theory...

The case of recommendation algorithms

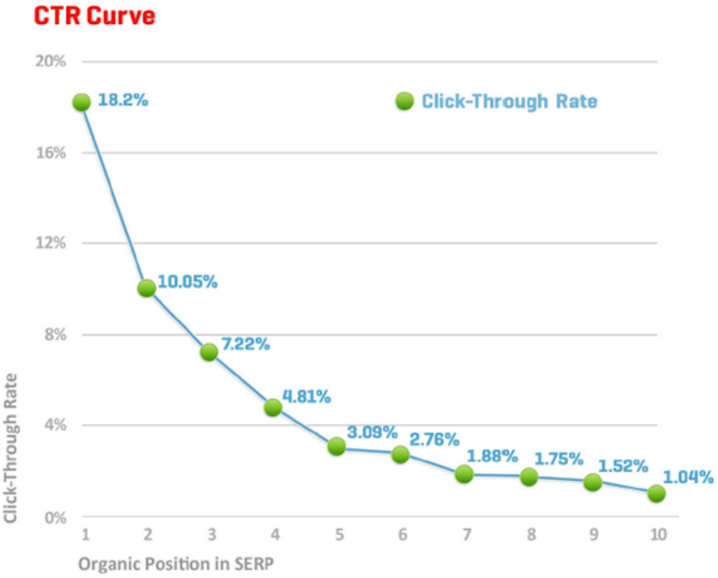
Recommenders

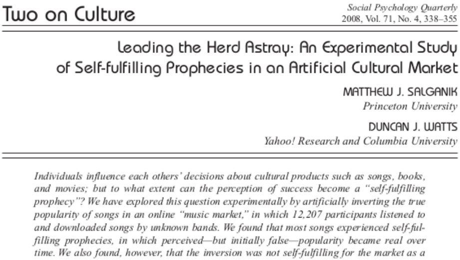
- Recommenders: filtering tools for items
- *Predict* user tastes for items
- Returns the most likely preferred items



...

Recommender impact





Two on Culture

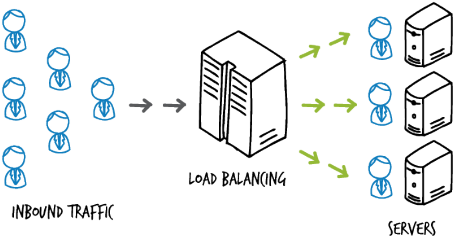
Social Psychology Quarterly
2008, Vol. 71, No. 4, 338–355

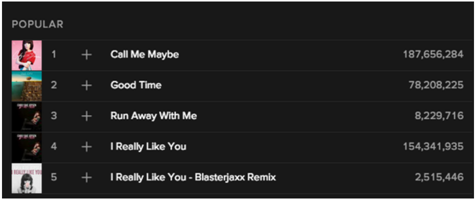
Leading the Herd Astray: An Experimental Study of Self-fulfilling Prophecies in an Artificial Cultural Market

MATTHEW J. SALGANIK
Princeton University

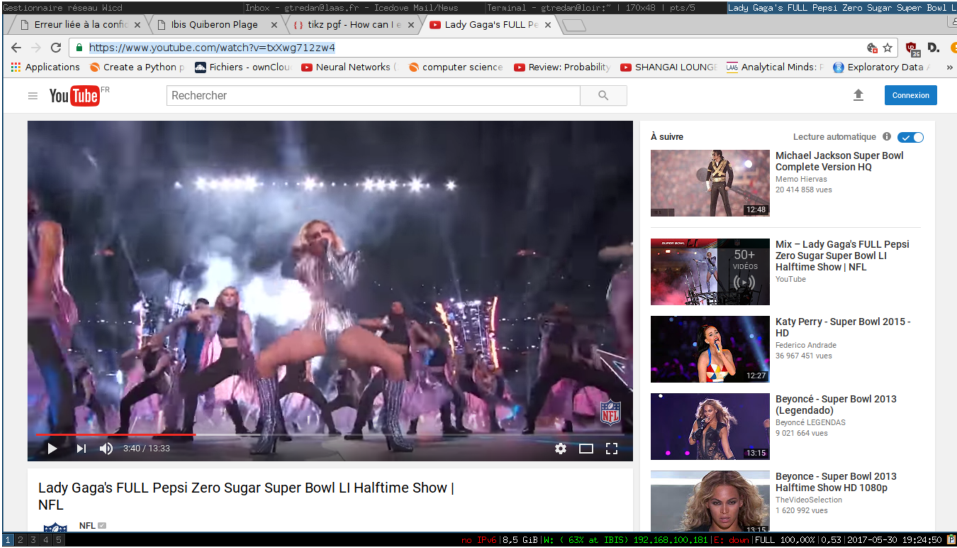
DUNCAN J. WATTS
Yahoo! Research and Columbia University

Individuals influence each others' decisions about cultural products such as songs, books, and movies; but to what extent can the perception of success become a "self-fulfilling prophecy"? We have explored this question experimentally by artificially inverting the true popularity of songs in an online "music market," in which 12,207 participants listened to and downloaded songs by unknown bands. We found that most songs experienced self-fulfilling prophecies, in which perceived—but initially false—popularity became real over time. We also found, however, that the inversion was not self-fulfilling for the market as a

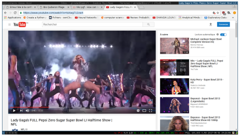




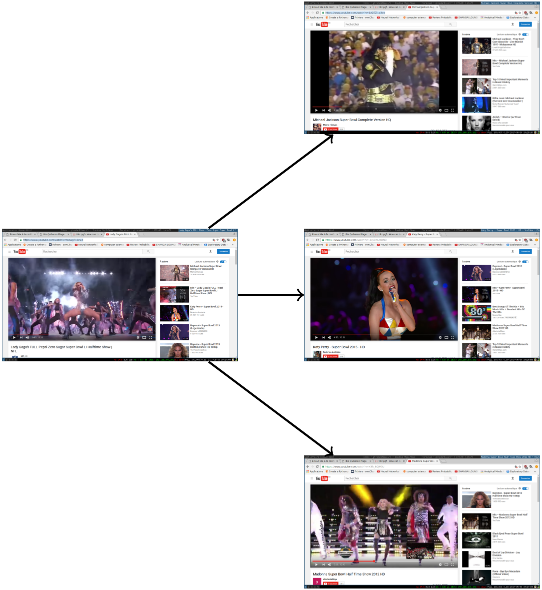
Crawling


Gestionnaire réseau Wicd
Inbox - gtredan@laas.fr - Icedove Mail/News
Terminal - gtredan@loir:~ | 170x48 | pts/5
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl L
Erreur liée à la confic
Ibis Quiberon Plage
tikz pgf - How can I e
Lady Gaga's FULL Pe
https://www.youtube.com/watch?v=txXwg712zw4
Applications
Create a Python p
Fichiers - ownClou
Neural Networks (
computer science
Review: Probability
SHANGAI LOUNGE
Analytical Minds: F
Exploratory Data
YouTube FR
Rechercher
Connexion
À suivre
Lecture automatique
Michael Jackson Super Bowl Complete Version HQ
Memo Hiervas
20 414 858 vues
12:48
Mix – Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
YouTube
50+ VIDÉOS
Katy Perry - Super Bowl 2015 - HD
Federico Andrade
36 967 451 vues
12:27
Beyoncé - Super Bowl 2013 (Legendado)
Beyoncé LEGENDAS
9 021 664 vues
13:15
Beyonce - Super Bowl 2013 Halftime Show HD 1080p
TheVideoSelection
1 620 992 vues
3:40 / 13:33
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
NFL
1 2 3 4 5
no IPv6 | 8,5 GiB | W: (63% at IBIS) 192.168.100.181 | E: down | FULL 100,00% | 0,53 | 2017-05-30 19:24:50

Crawling

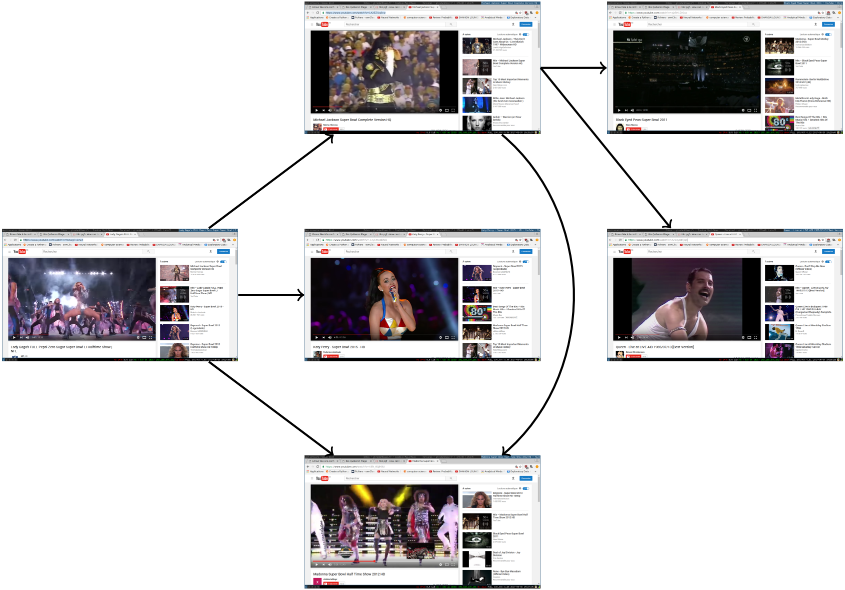

Rechercher
À suivre
Michael Jackson Super Bowl Complete Version HQ
Mix – Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
Katy Perry - Super Bowl 2015 - HD
Beyoncé - Super Bowl 2013 (Legendado)
Beyonce - Super Bowl 2013 Halftime Show HD 1080p
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL

Crawling

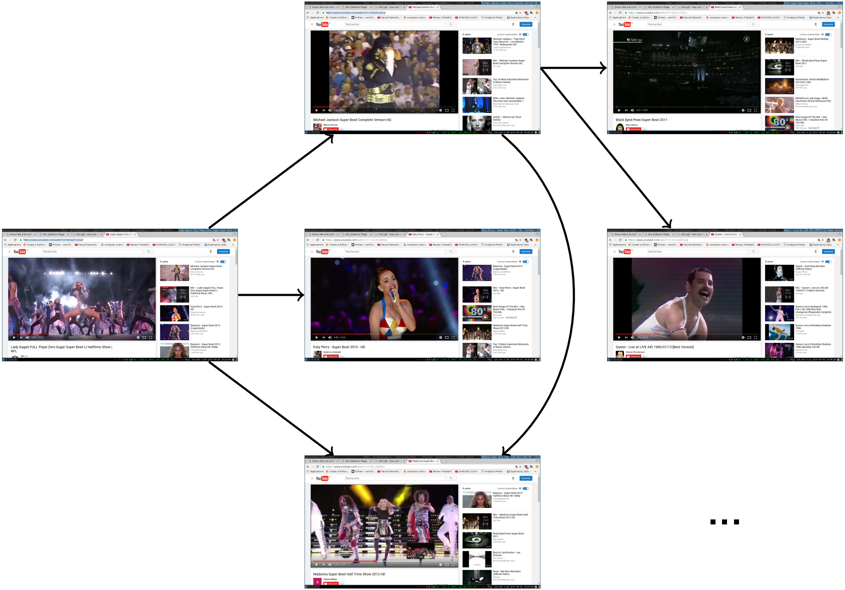


Michael Jackson Super Bowl Complete Version HQ
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
Katy Perry - Super Bowl 2015 - HD
Madonna Super Bowl Half Time Show 2012 HD

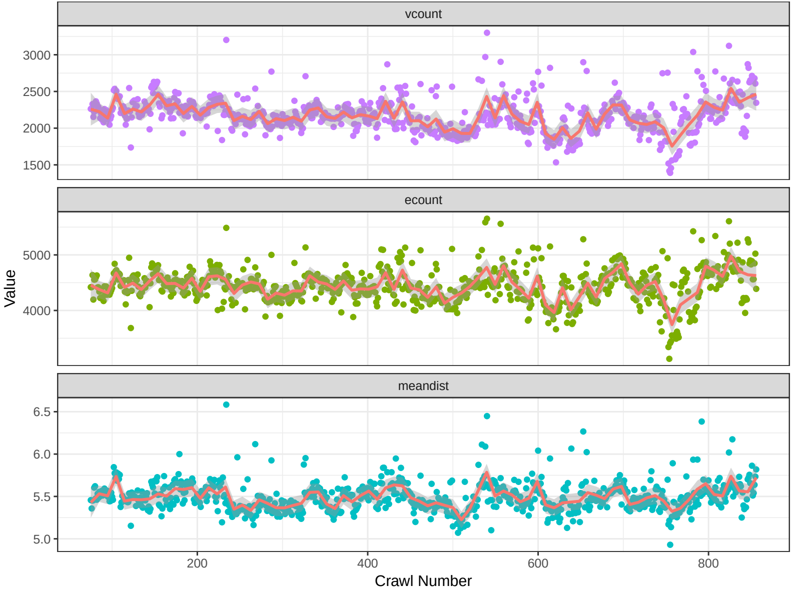
Crawling


Michael Jackson Super Bowl Complete Version HQ
Black Eyed Peas-Super Bowl 2011
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
Katy Perry - Super Bowl 2015 - HD
Queen - Live at LIVE AID 1985/07/13 [Best Version]
Madonna Super Bowl Half Time Show 2012 HD

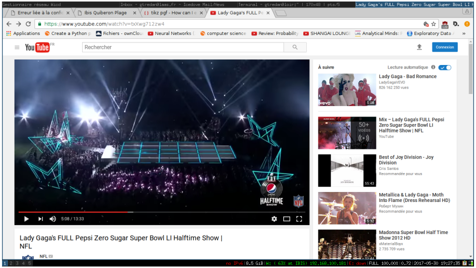
Crawling



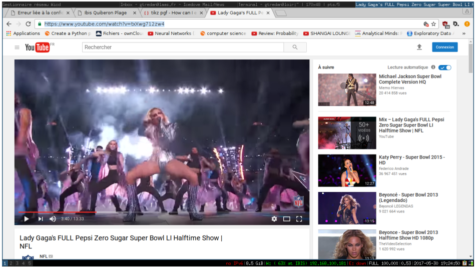
Michael Jackson Super Bowl Complete Version HQ
Black Eyed Peas-Super Bowl 2011
Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
Katy Perry - Super Bowl 2015 - HD
Queen - Live at LIVE AID 1985/07/13 [Best Version]
Madonna Super Bowl Half Time Show 2012 HD
...



33 days of youtube hourly crawling. $\mathbb{E}(|V(G_t) \cap V(G_{t+1})|) = 74.7\%$

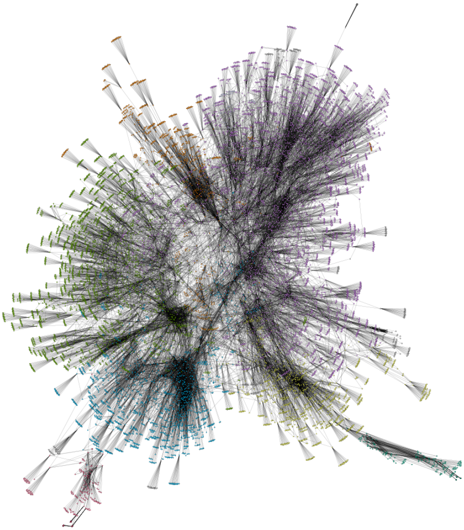


With cookie



Without cookie

With cookie



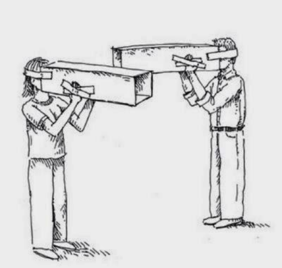
Without cookie

Bias

What is bias ?

Difficult to define

- Political (soft censorship)
- Economical (maximise income)
- Operational (serendipity)



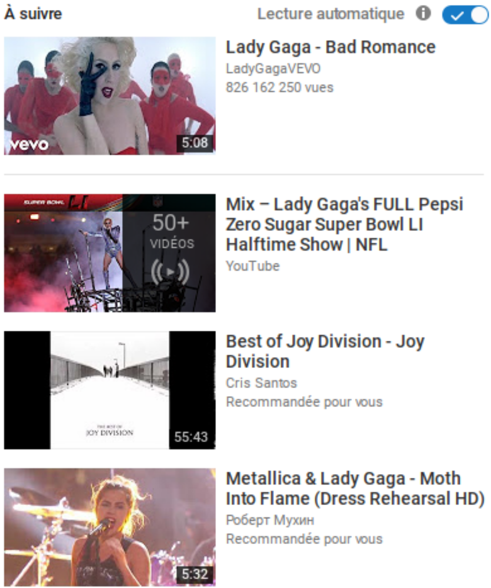
Our definition:

Biasing edges= rewiring the graph of recommendations

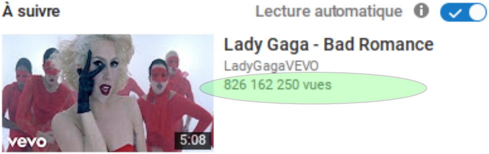
Observation 1 Biased edges *tangibly* impact the graph structure

Observation 2 It is possible to detect such bias.

Dataset

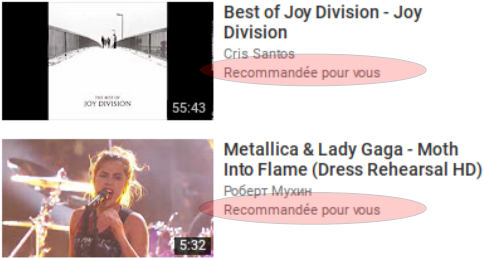

À suivre
Lecture automatique
Lady Gaga - Bad Romance
LadyGagaVEVO
826 162 250 vues
vevo
5:08
Mix – Lady Gaga's FULL Pepsi Zero Sugar Super Bowl LI Halftime Show | NFL
YouTube
50+
VIDÉOS
Best of Joy Division - Joy Division
Cris Santos
Recommandée pour vous
JOY DIVISION
55:43
Metallica & Lady Gaga - Moth Into Flame (Dress Rehearsal HD)
Роберт Мухин
Recommandée pour vous
5:32

Dataset

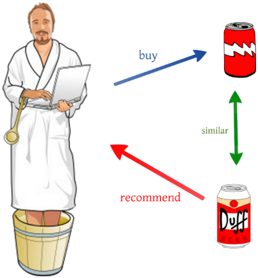


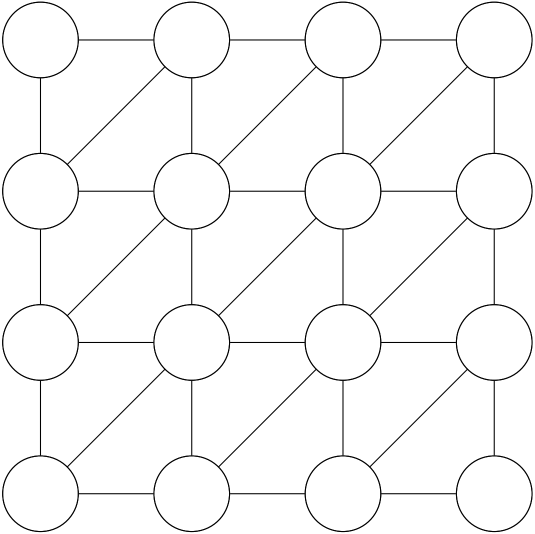
$k = 17$ normal recommendations

$k' = 2$ "Recommended for you"



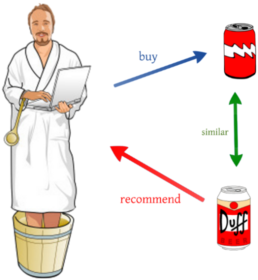
Analogy: Locality model



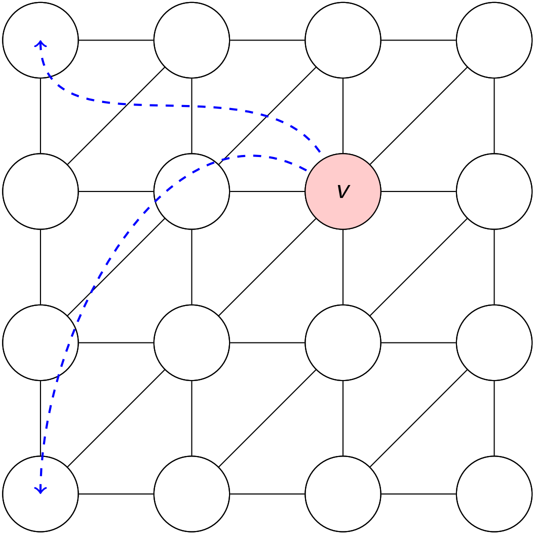


- Short links $\leftrightarrow$ "locality" "Homophily"

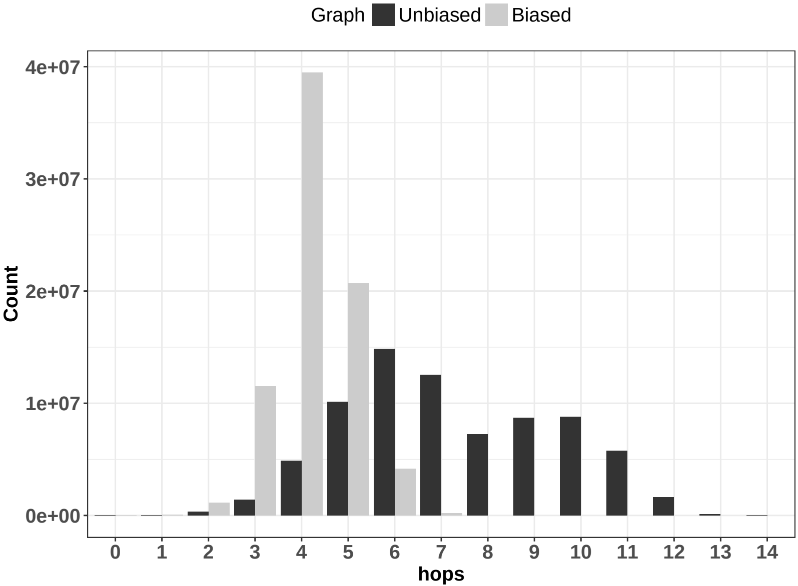
Analogy: Locality model



- Short links ↔ "locality" "Homophily"
- Long "random" links ↔ weak ties

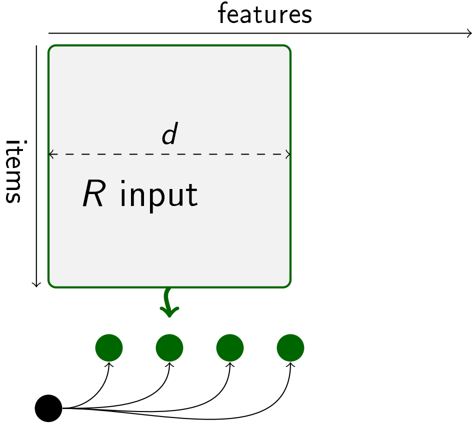


Distance impact



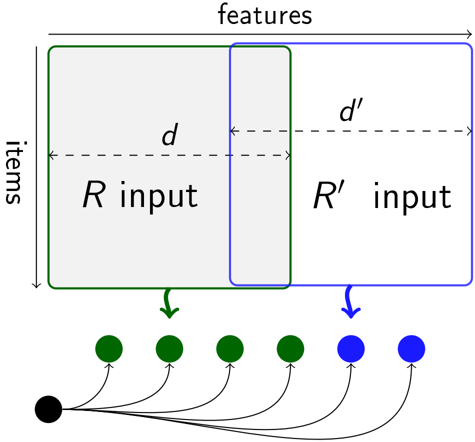
A Toy Model

Objective: tune the level of bias introduced by the operator



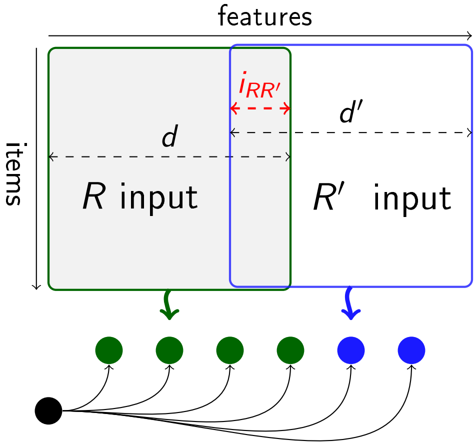
A Toy Model

Objective: tune the level of bias introduced by the operator



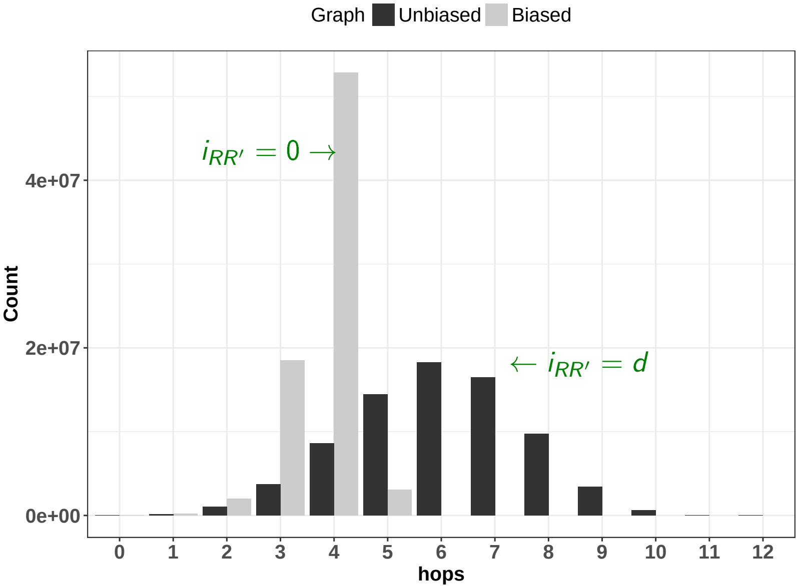
A Toy Model

Objective: tune the level of bias introduced by the operator



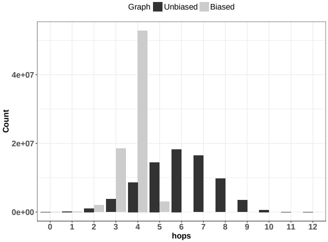
- $i_{RR'} = 0$: Independent outputs, "maximum bias"
- $i_{RR'} = d = d'$: No bias

Distance impact

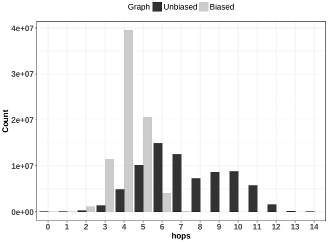


Detecting biased edges

Detection - Approach



Toy model

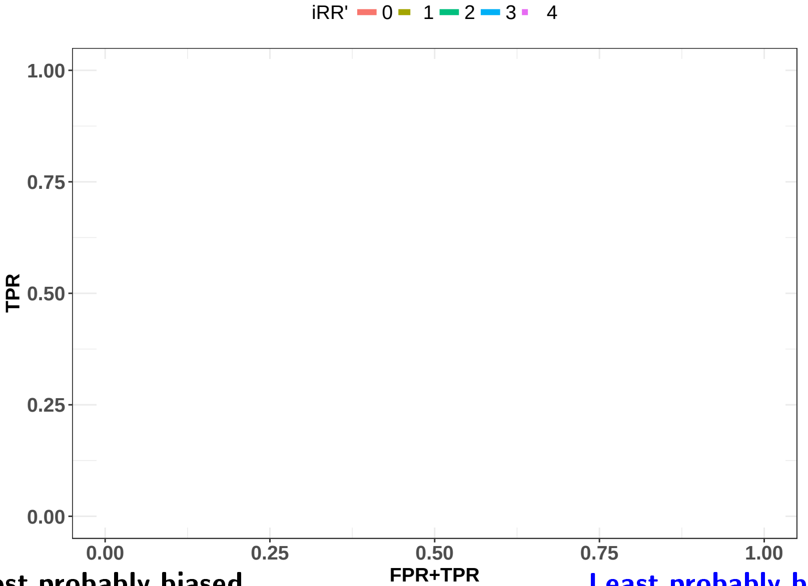


YT dataset

- The removal of 10% links has a drastic impact on path length distribution
- $\Rightarrow$ **important** links (wrt hop distance)
- $\Rightarrow$ Betweenness centrality should do:

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \propto \mathbb{P}(e \in \mathit{Biased})$$

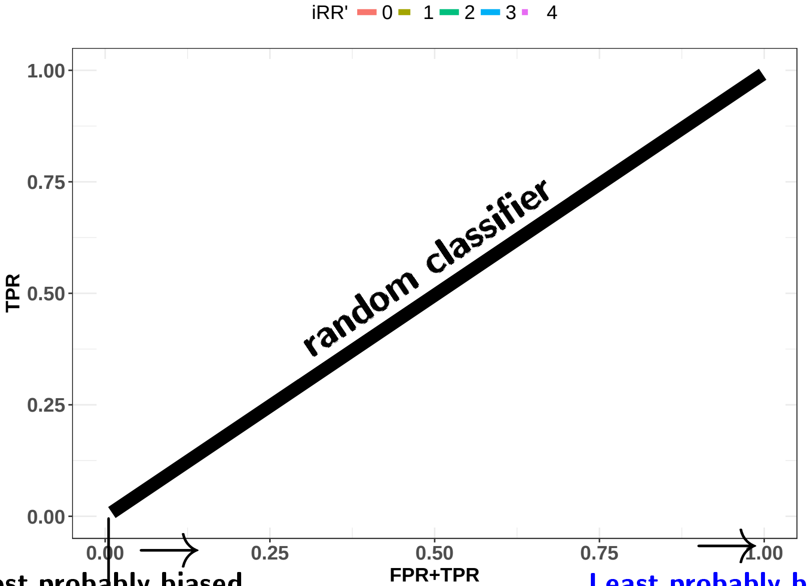
Detection - Model



Most probably biased

Least probably biased

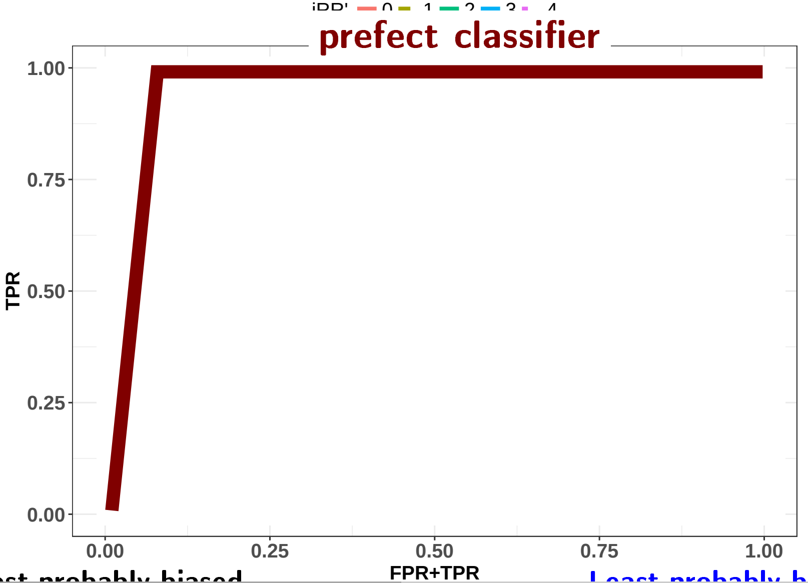
Detection - Model



Most probably biased

Least probably biased

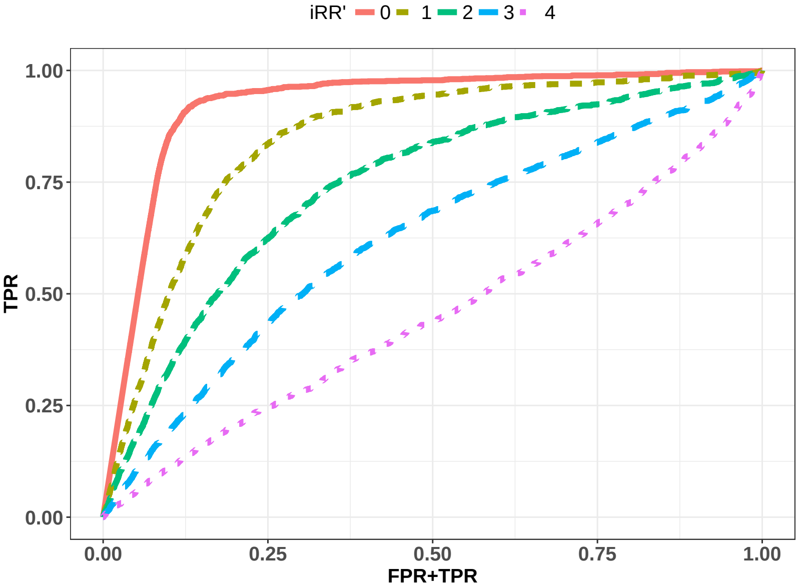
Detection - Model



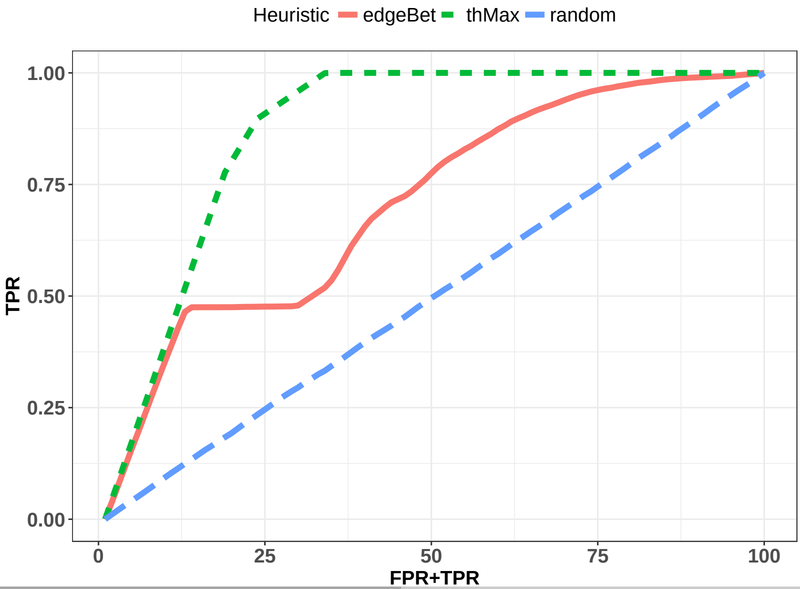
Most probably biased

Least probably biased

Detection - Model



Detection - Youtube



Conclusion

- Example application: bias detection
  - Bias "breaks" the recommender locality
  - Not so bad heuristic
  - User-local observation !

  The topological face of recommendation, Complex Networks, 2017.

- "Reverse engineering" remote black boxes
- ... Difficult model but...
- only answers to a few questions


Couverture des publications
Le nombre de personnes pour qui des publications de votre Page se sont affichées sur leur écran. Ce chiffre est une
être précis.
Organique
Payé
600K
500K
400K
300K
200K
100K
0
10
12
14
16
18
20
22
24
26
28
30
03
05
AOÛ
SEP
Couverture des publications
Le nombre de personnes pour qui des publications de votre Page se sont affichées sur leur écran. Ce chiffre est
pas être précis.
Organique
Payé
Dé-référencement


Canon


Canon