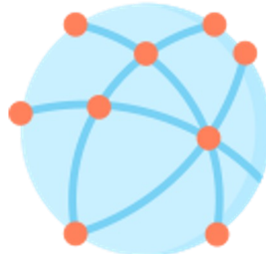


Assessing Research Impact by Leveraging Open Scholarly Knowledge Graphs

The Web Conference 2022 - Tutorials Session

Ilias Kanellos

Dimitris Sacharidis

Thanasis Vergoulis

Part A: Introduction

Dimitris Sacharidis (Université Libre de Bruxelles, Belgium)

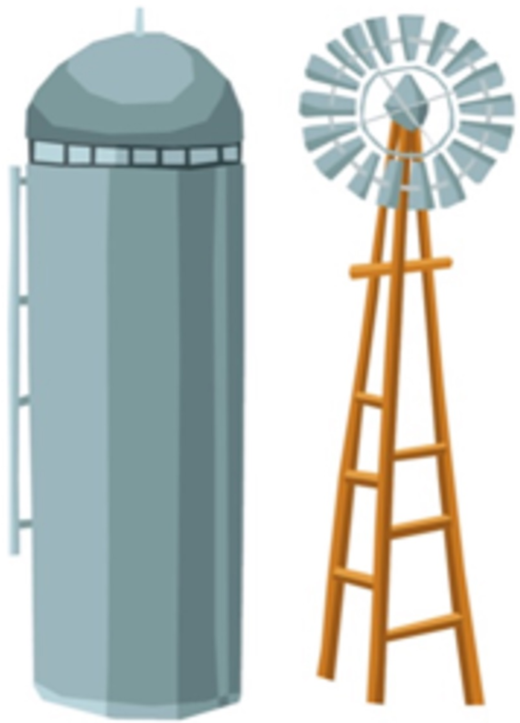
Standing on the Shoulders of Giants

- ***Scholarly communication*** is paramount to advancing science.
- How to find the ***most valuable publications***?
- Two problems:
  - ***Discovery***
  - ***Impact Assessment***

https://www.vecteezy.com/free-vector/business

Discovery of Scholarly Knowledge

- ***Publisher silos:*** publishers control the dissemination of scientific articles



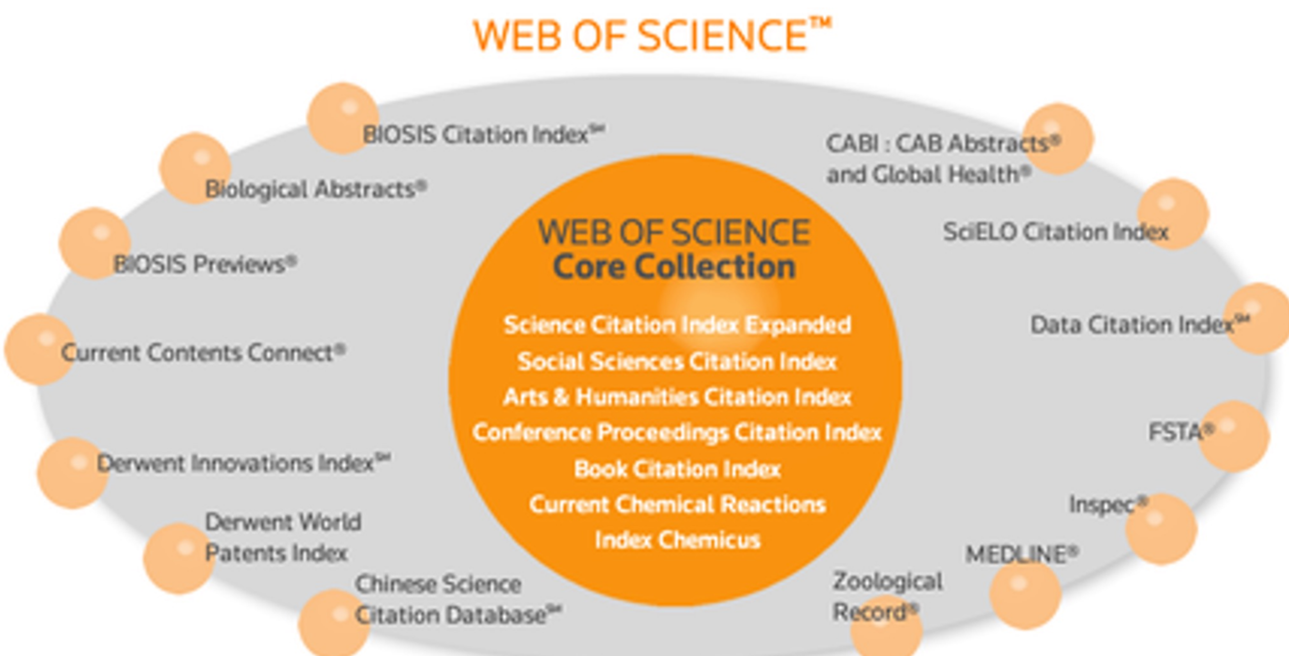
https://www.freepik.com/free-vector/farm-decorative-multicolored-set_3977275.htm

Discovery options:

- Directly from the **publisher**
- *From **citation indices***
- Using Web ***search engines***

Citation Indices: Web of Science

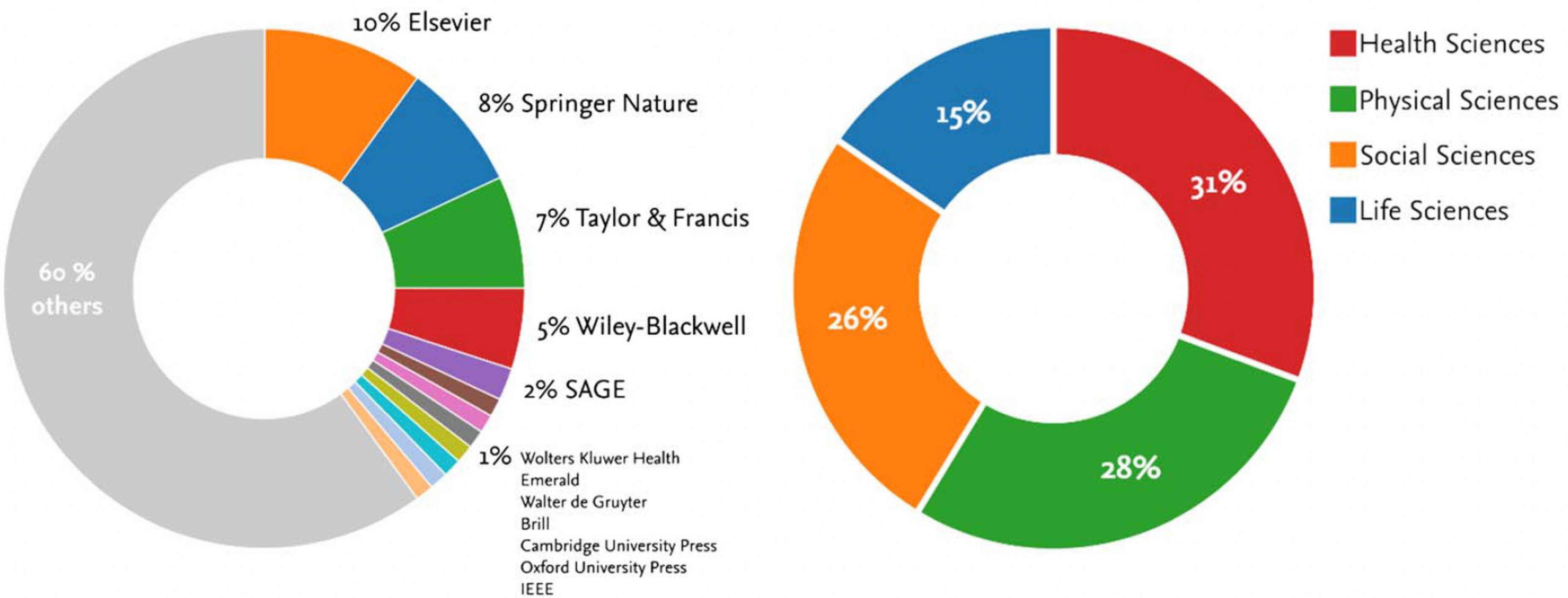
- ***Web of Science*** from Clarivate Analytics
  - Based on the Science Citation Index founded by Eugene Garfield in 1964



By Thomson Reuters - The relevant Thomson Reuters web page., Fair use, https://en.wikipedia.org/w/index.php?curid=44053851

Citation Indices: Scopus

- ***Scopus*** from Elsevier



https://doi.org/10.1162/qss_a_00019

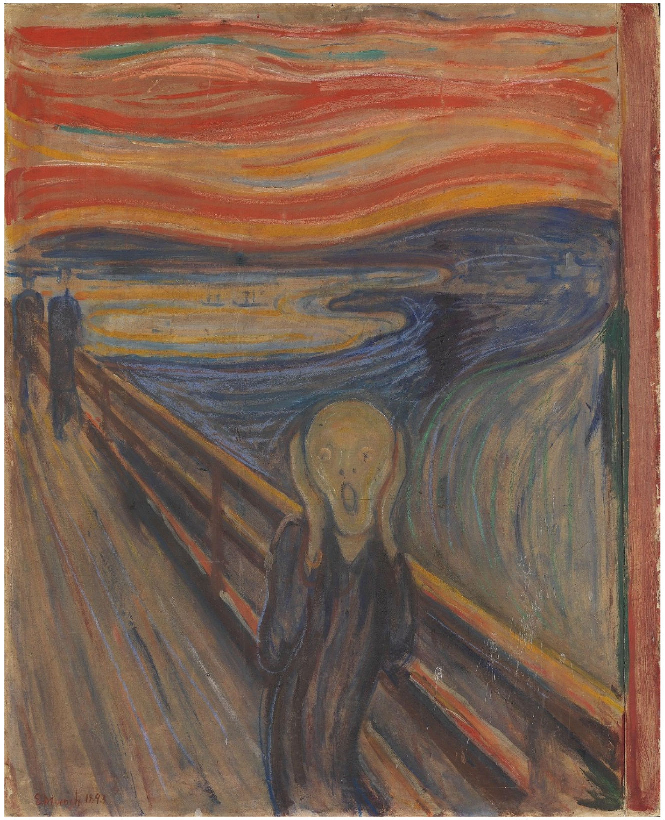
Published Research is Exponentially Increasing

- The ***growth rate*** of the number of published research is ***constantly increasing***.
- Studies suggest that, among the vast number of published works, many are of ***questionable quality or low impact***.
- Identifying most ***valuable publications*** for any given research topic has become ***tedious & time consuming***.

Photo by Carles Rabada on Unsplash

Why?

- ***Increase in the number of researchers worldwide***.
    - ^20% between 2007-2014*
- ***Publish or Perish***
    - incredible pressure to publish more, especially on young researchers



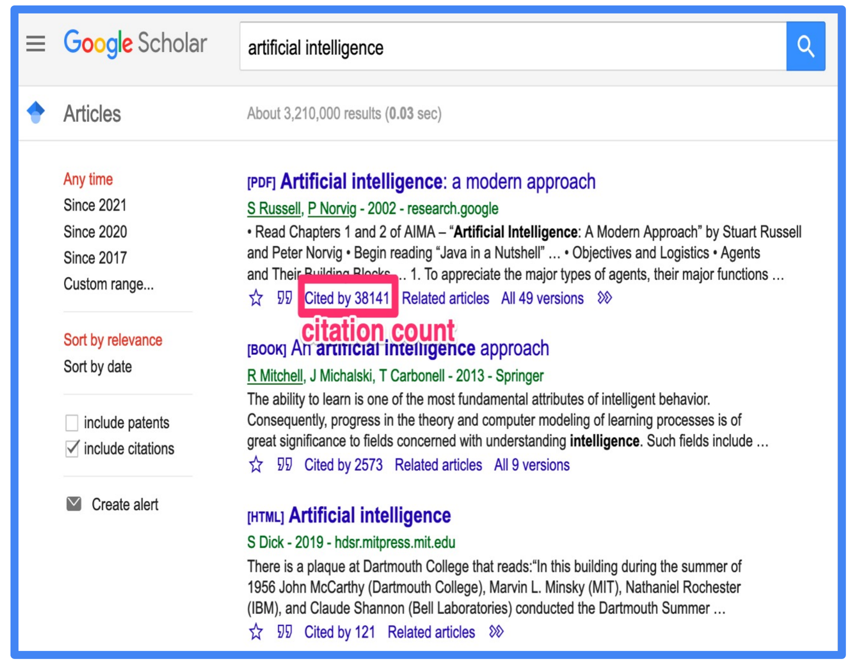
Edvard Munch, "The scream of nature"
https://bit.ly/3dcLbXD

*UNESCO Science Report: towards 2030. UNESCO Publishing, 2015

Impact Assessment

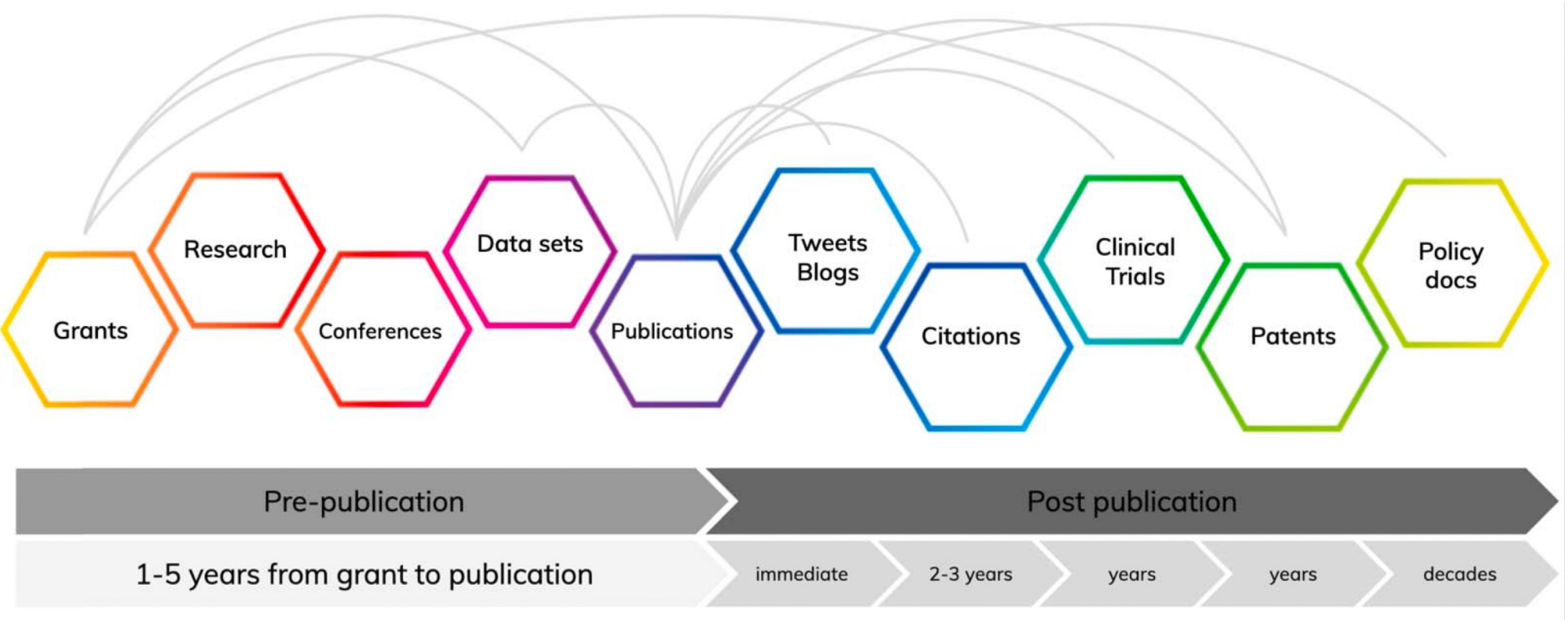
- ***Quantifying the impact*** of publications could facilitate the identification of valuable research.
  - ***Open Science initiatives***, having momentum make the calculation of such measures possible.

- Academic search engines ***combine keyword-search with a scientific impact measure*** (usually citation counts) to rank publications.
  - possible other applications



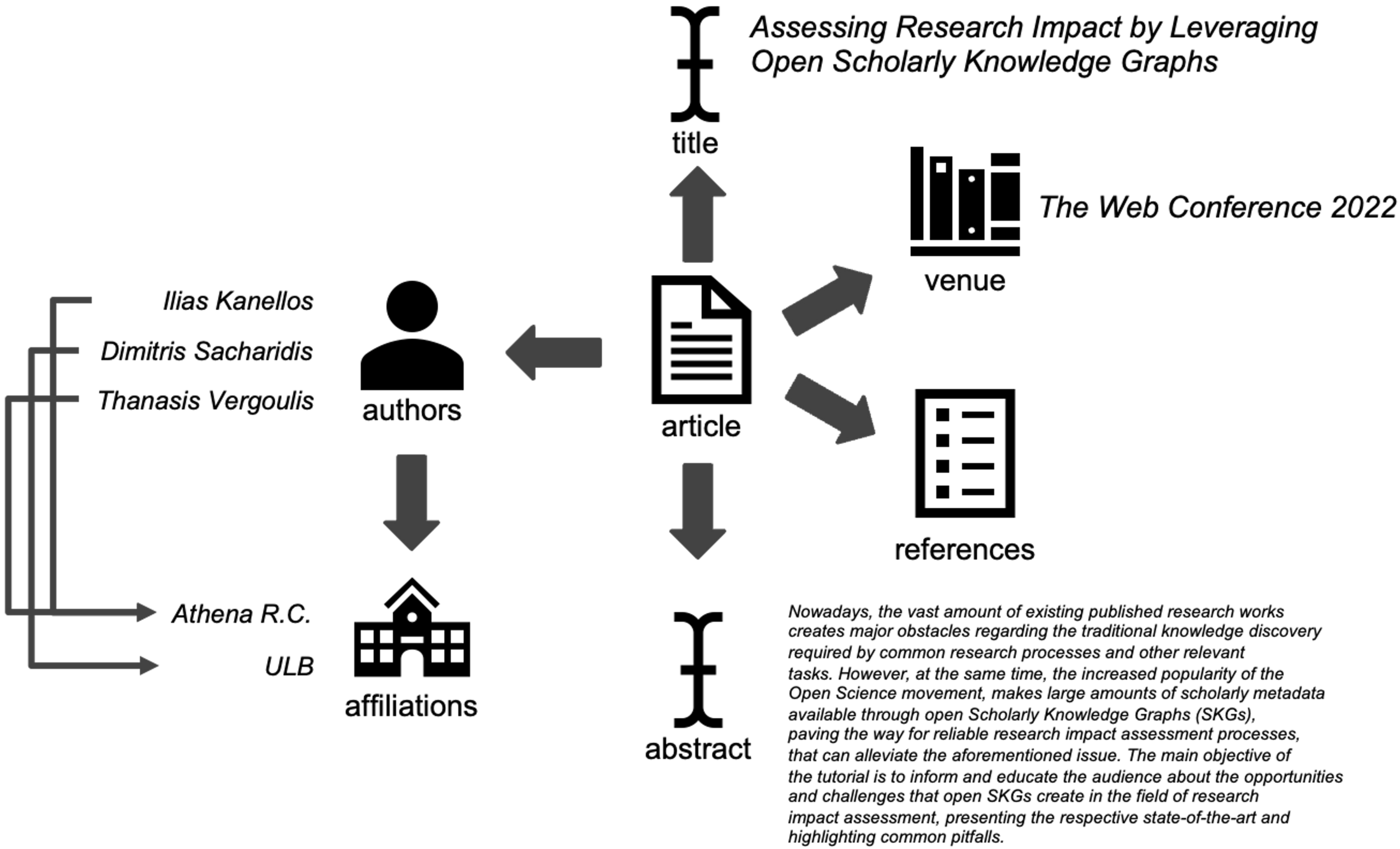
From what data?

Scholarly Communication Lifecycle

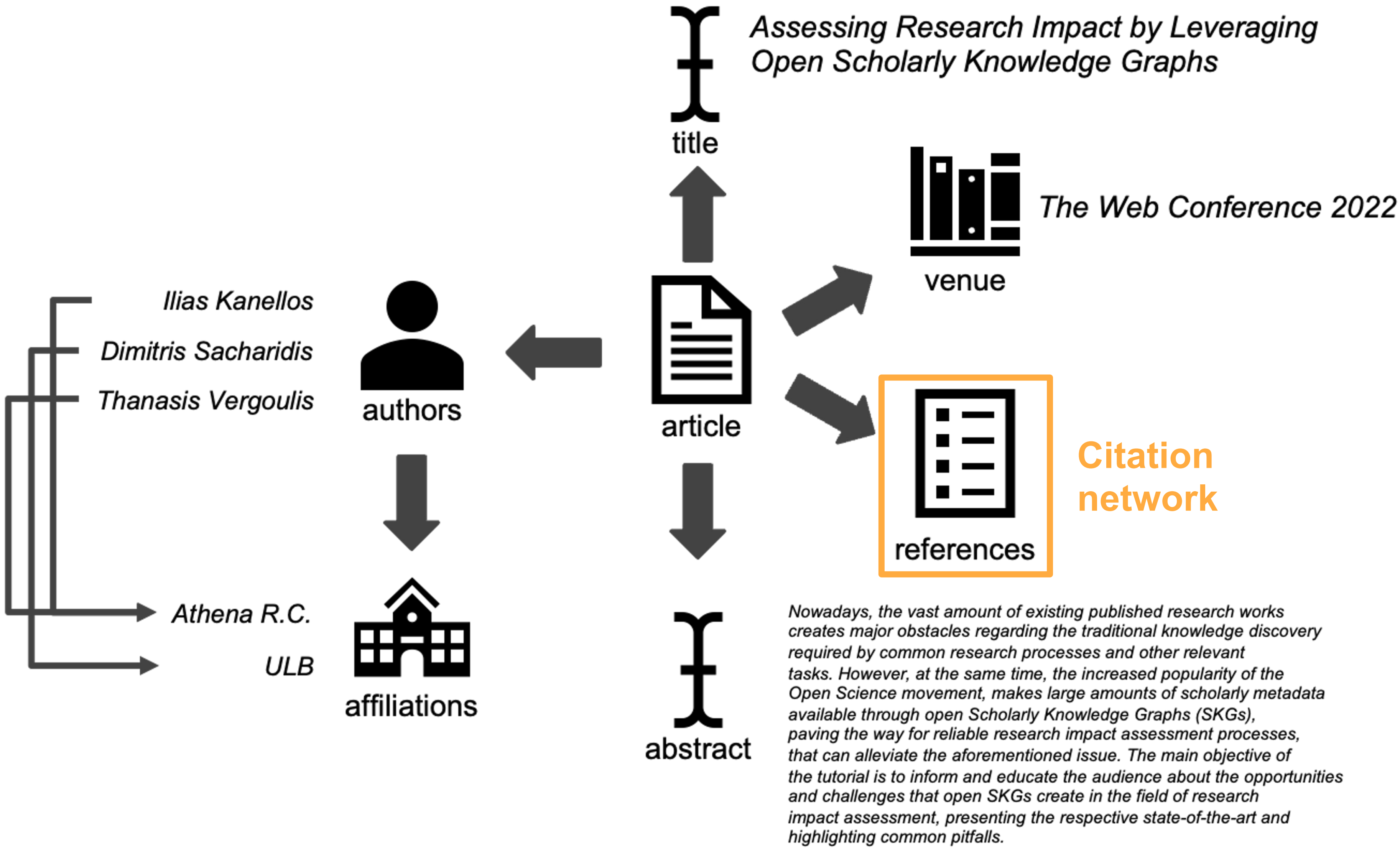


https://doi.org/10.1162/qss_a_00020

Scholarly Communication Metadata



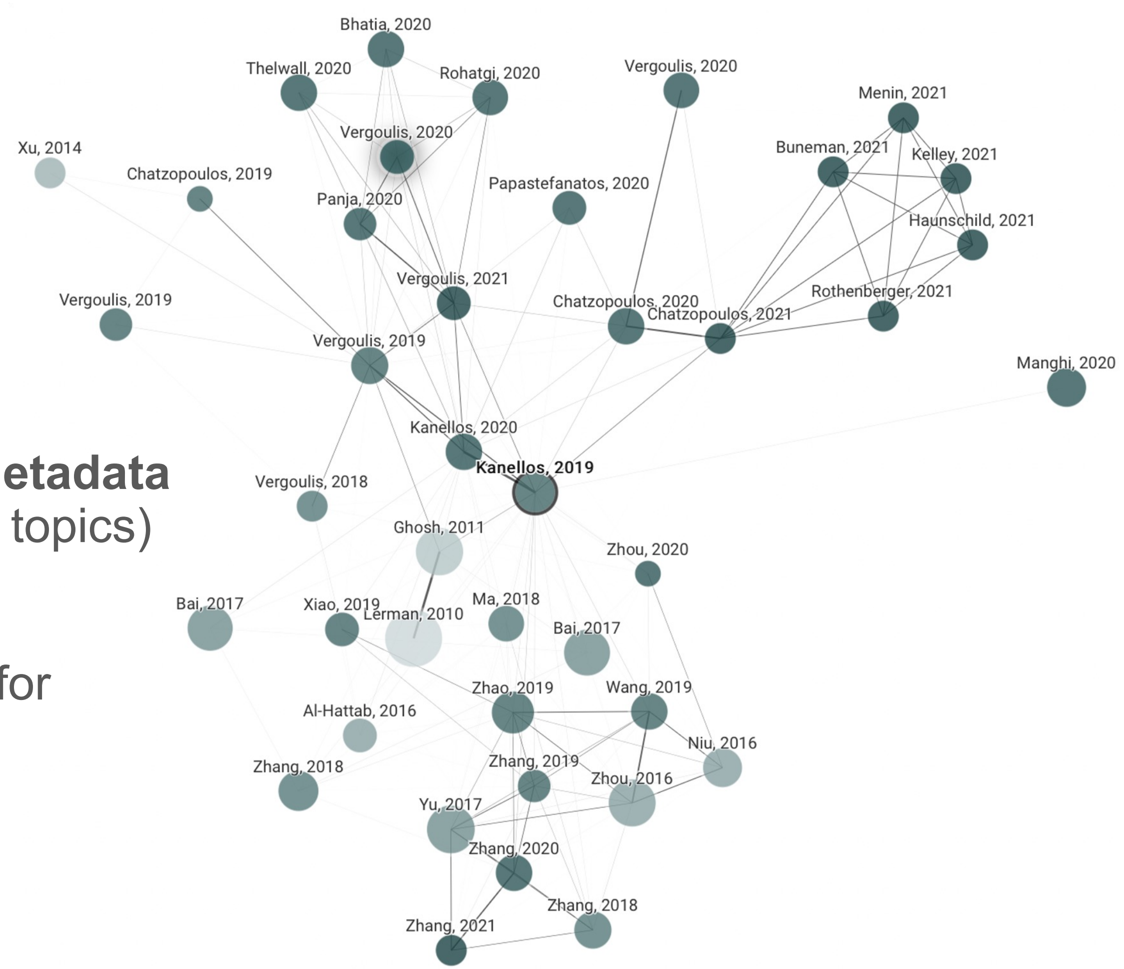
Scholarly Communication Metadata



Citation Networks

Connect **articles** and their **metadata** (authors, affiliations, venues, topics)

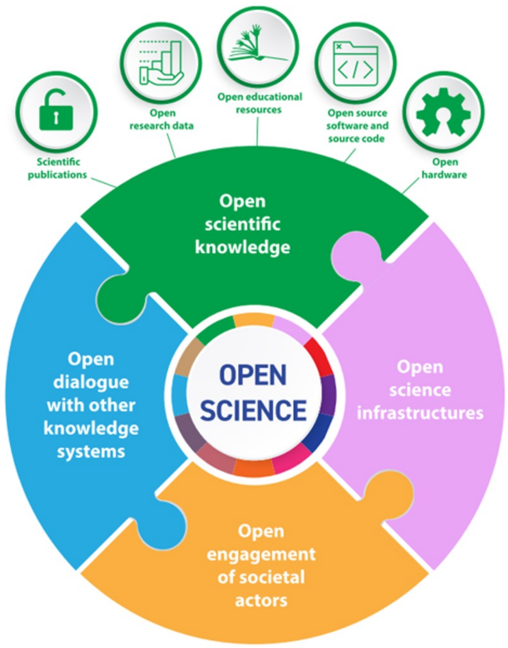
Capture useful relationships for **impact assessment**



Where's the Metadata?

Behind publisher's silos, or publisher-driven ***paid*** citation indexes.

https://www.freepik.com/free-vector/farm-decorative-multicolored-set_3977275.htm



https://www.openscience-twente.com/open-science/#open-scientific-knowledg

But recent ***Open Science*** initiatives help make the metadata publicly available

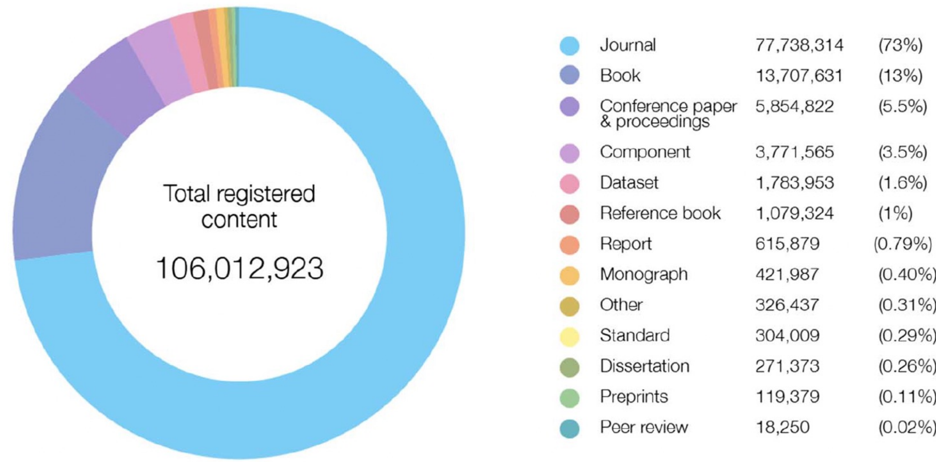
I4OC

Initiative for Open Citations

Crossref

In 1999, publishers agree to use ***Digital Object Identifiers*** (DOIs) to link their articles

Crossref was born as a not-for-profit ***association*** having publishers as **members**, and allowing them to register their **DOIs**.

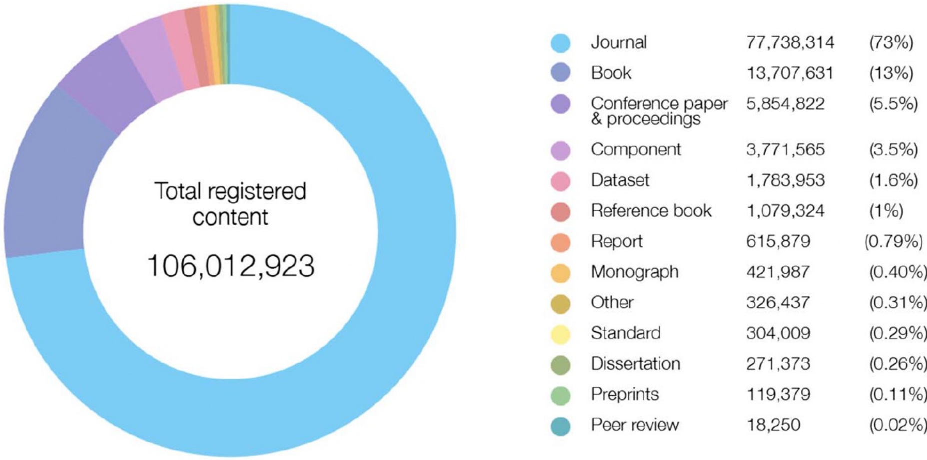


Reference setting per DOI prefix	What this means for reference distribution
Closed	These references are only used for the Crossref Cited-by service (members-to-members) and are not distributed via any of the public interfaces or APIs.
Limited	In addition, organizations that sign an agreement for Crossref's Metadata "Plus" subscription-based service can access these references. (This is the default for older membership accounts pre-2017.)
Open	Everyone can access these references through our open APIs. (This is the default for accounts joining from 2018.)

Crossref

In 1999, publishers agree to use ***Digital Object Identifiers*** (DOIs) to link their articles

Crossref was born as a not-for-profit ***association*** having publishers as **members**, and allowing them to register their **DOIs**.



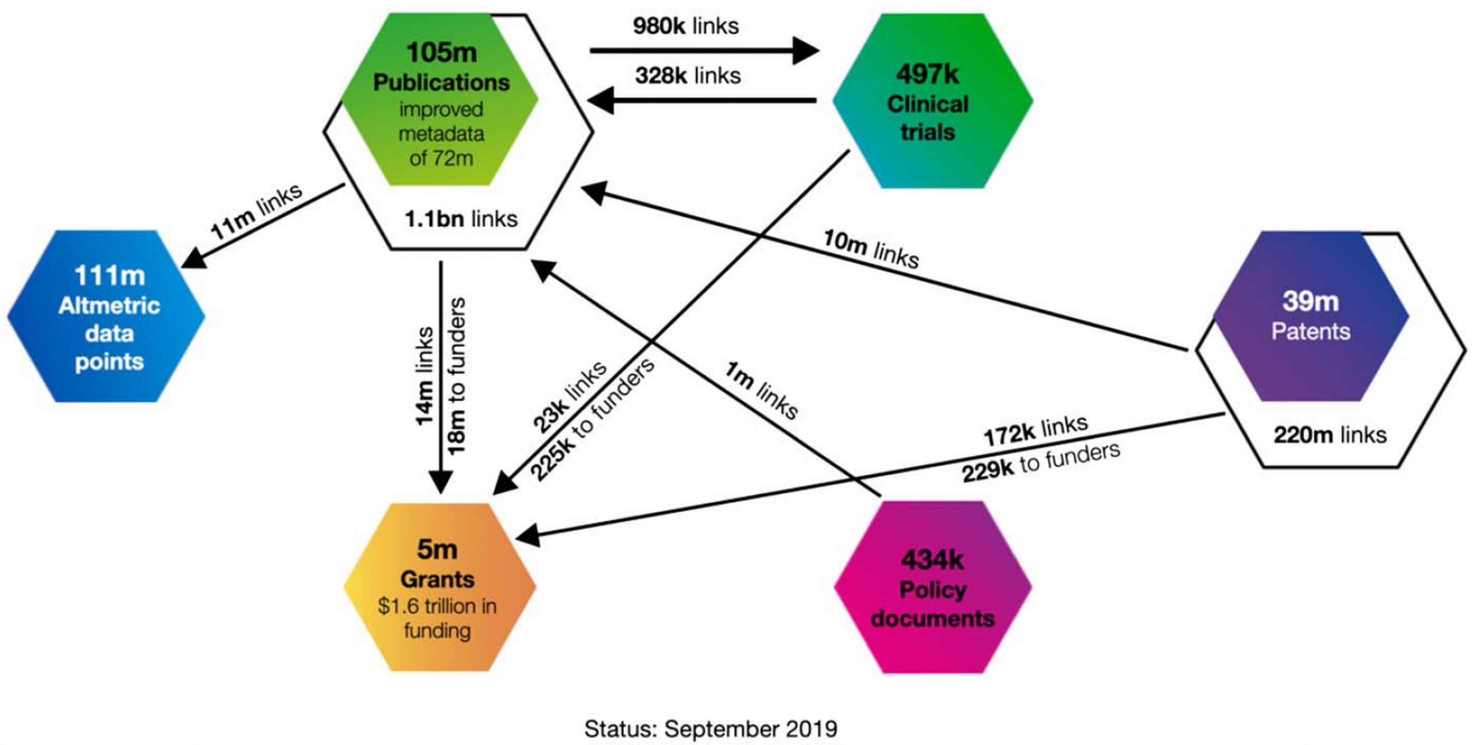
Reference setting per DOI prefix	What this means for reference distribution
Closed	These references are only used for the Crossref Cited-by service (members-to-members) and are not distributed via any of the public interfaces or APIs.
Limited	In addition, organizations that sign an agreement for Crossref's Metadata "Plus" subscription-based service can access these references. (This is the default for older membership accounts pre-2017.)
Open	Everyone can access these references through our open APIs. (This is the default for accounts joining from 2018.)

I4OC

Dimensions

Dimensions from Digital Science collects ***rich metadata*** using Crossref as the backbone

- Offers a free-tier for accessing their data for scientific purposes

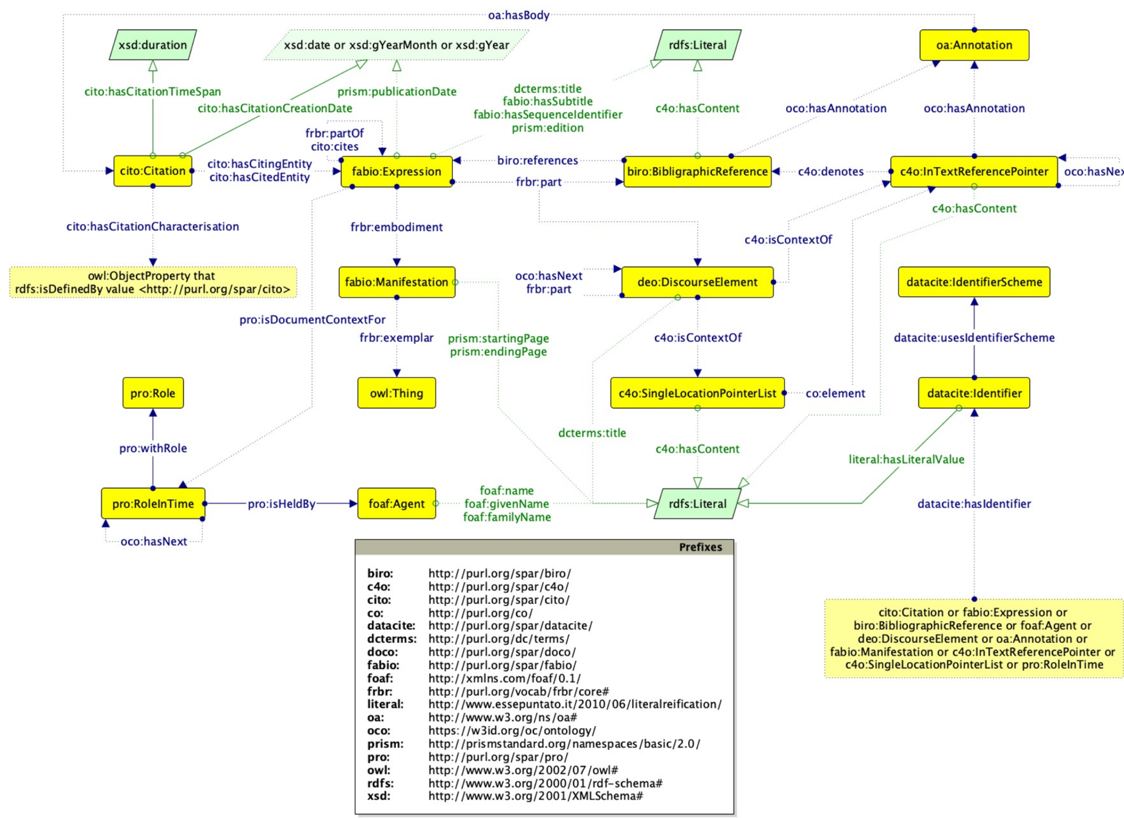


https://doi.org/10.1162/qss_a_00020

Open Scholarly Knowledge Graphs

Use ***Linked Open Data*** technologies to provide access to scholarly metadata

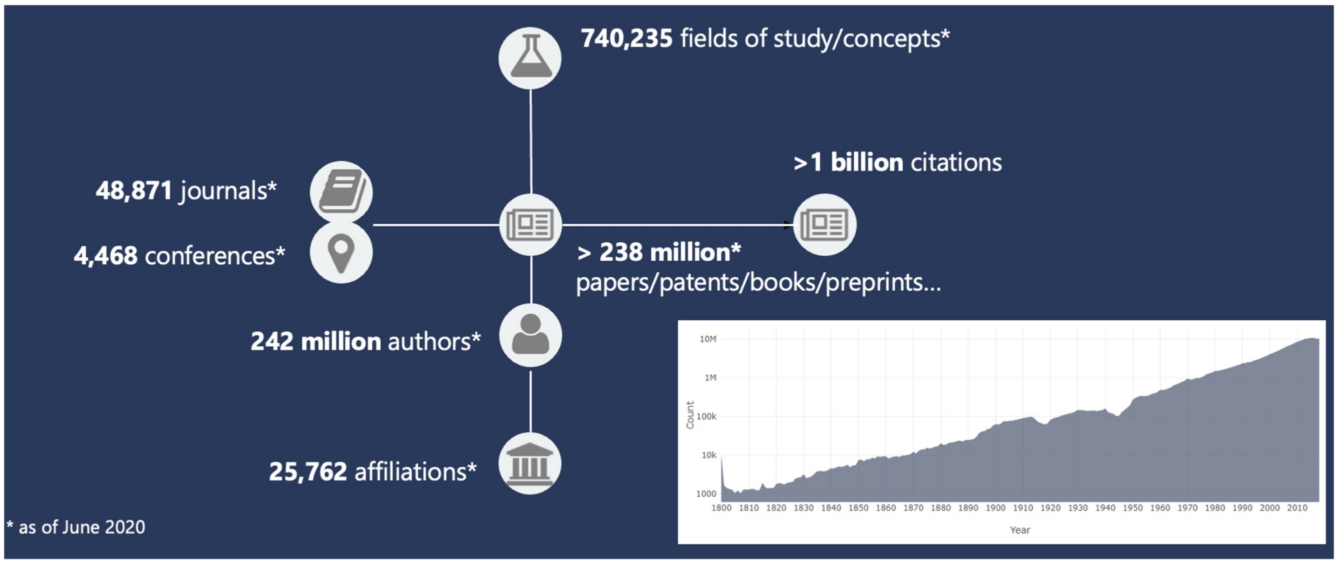
OpenCitations publishes the Index of Crossref Open DOI-to-DOI Citations (COCI)



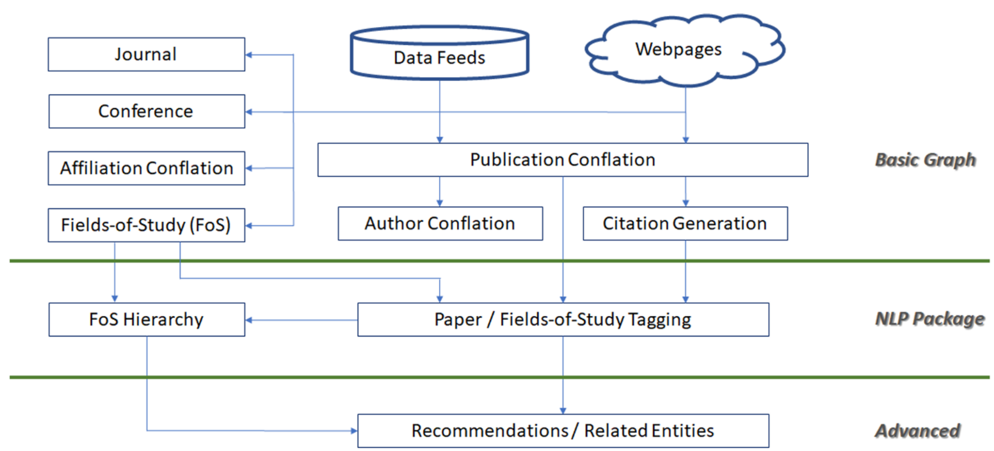
https://opencitations.net/model

Microsoft Academic Graph

The data powering **Microsoft Academic Services**



uses AI/ML technologies to parse **Web content**



Microsoft Academic, Alvin Chen, Chieh-Han Wu

Microsoft Academic Graph

Is now **retired**

Microsoft Academic / Blog

Next Steps for Microsoft Academic – Expanding into New Horizons

May 4, 2021

- Microsoft Academic Graph/Microsoft Academic Knowledge Exploration Service: No longer providing updated data or access to old releases after Dec. 31, 2021; however, existing copies can still be used under license.

But **alternatives** are emerging

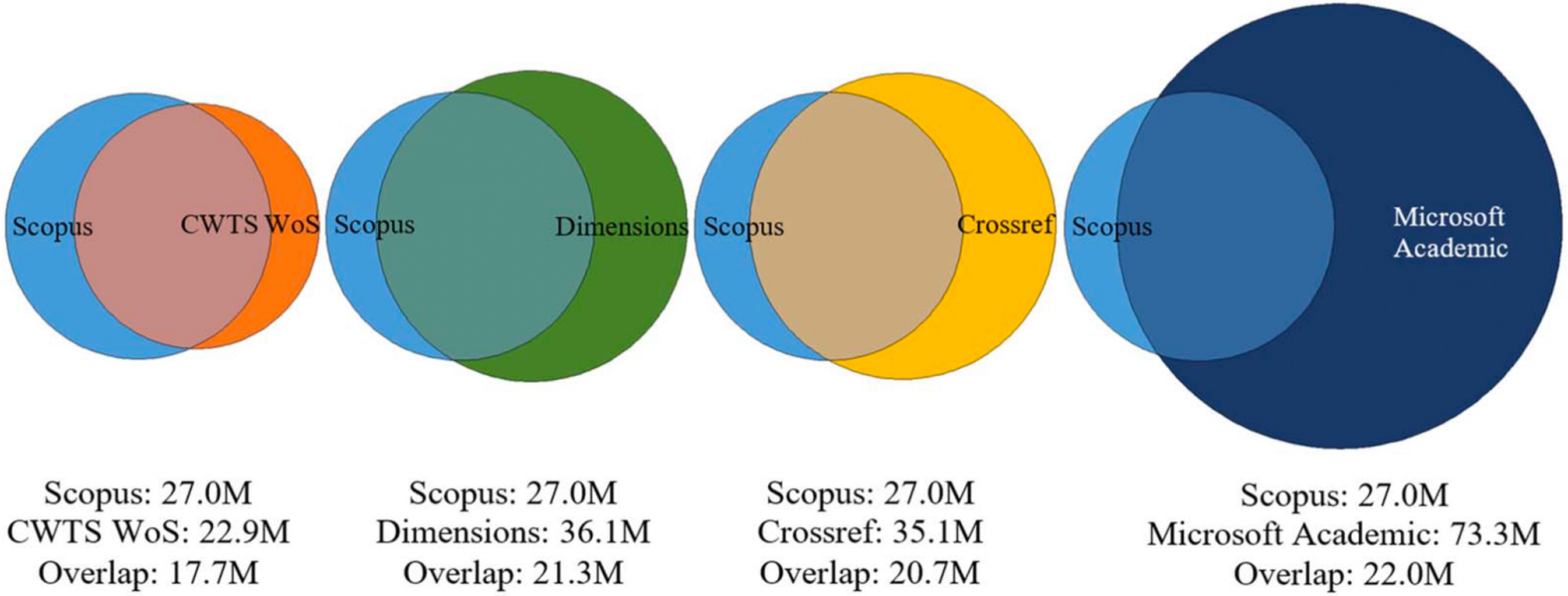
OpenAlex

An open and comprehensive catalog of scholarly papers, authors, institutions, and more.

Inspired by the ancient Library of Alexandria, OpenAlex is an index of hundreds of millions of interconnected entities across the global research system. We're 100% free and open source, and offer access via a web interface, API, and database snapshot.

How Good is the Data?

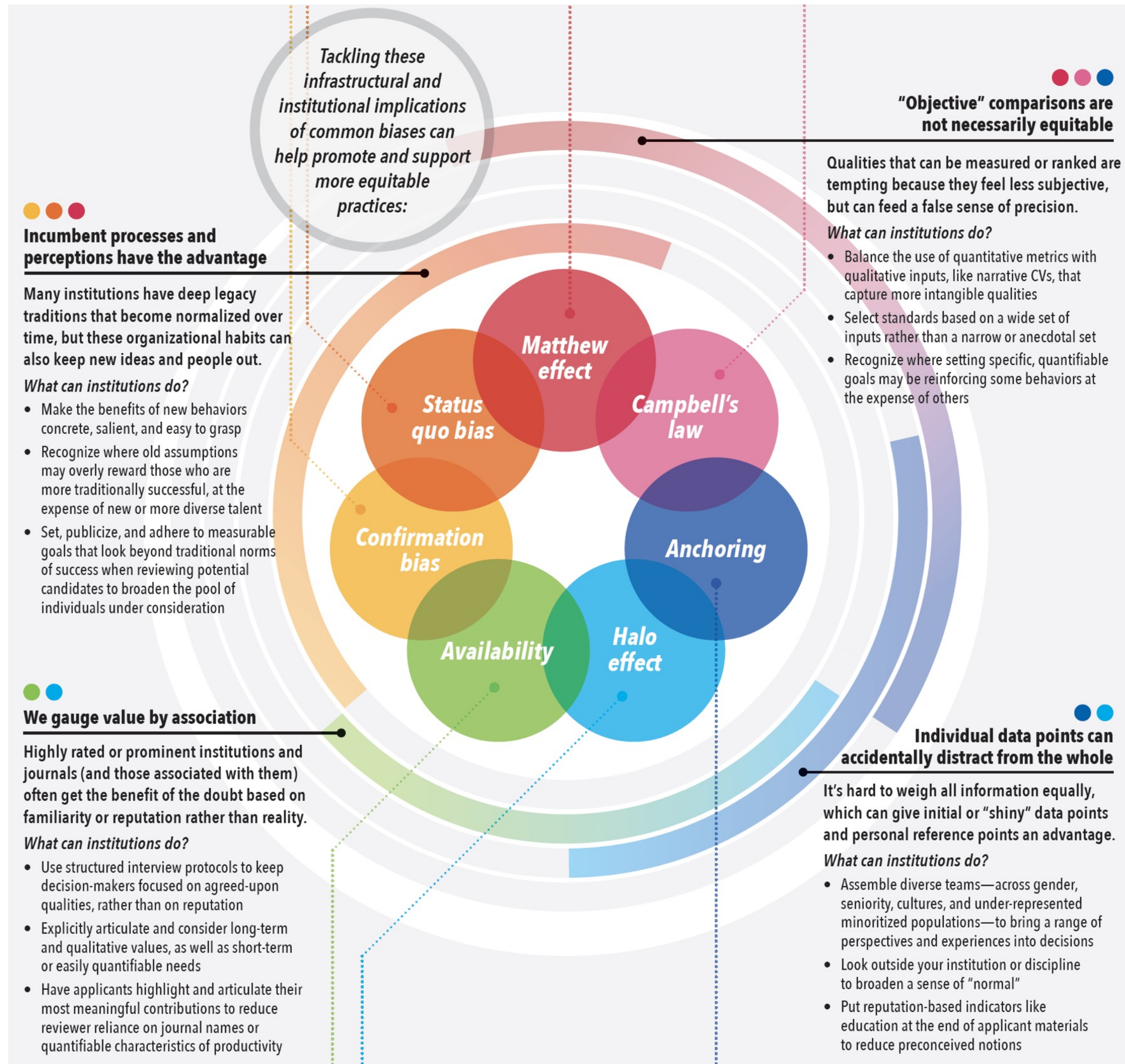
Metadata providers have different focus, sources, tools, and may differ greatly in ***coverage*** and ***quality***



https://doi.org/10.1162/qss_a_00112

How do you Assess Impact?

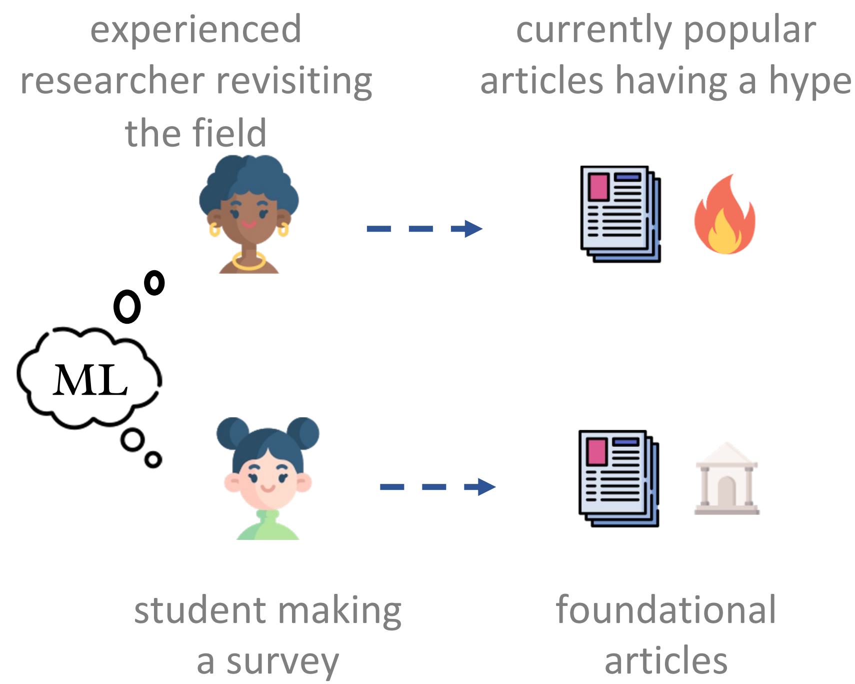
Even if the metadata is of sufficient quality, there are ***biases*** and ***pitfalls*** to consider when assessing impact



https://sfdora.org/resource/rethinking-research-assessment-unintended-cognitive-and-systems-biases/

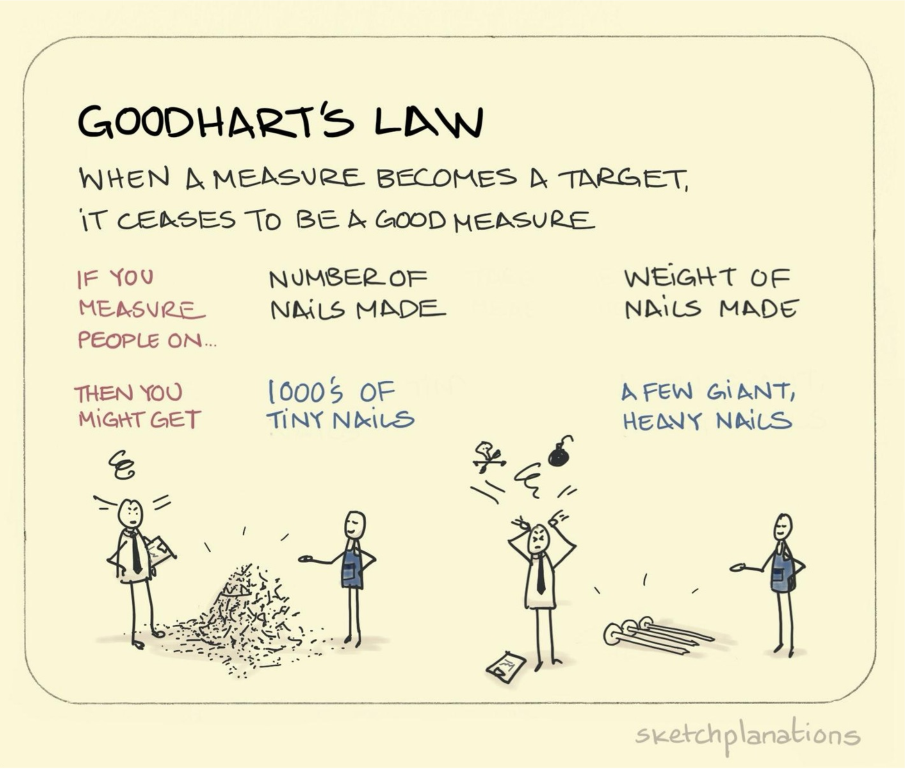
Pitfall #1: Scientific impact has various aspects

- It is an ***oversimplification to rely only on one impact measure***, like most academic search engines.
  - There are many ***diverse aspects*** of scientific impact, each most appropriate in different scenarios.
- Also there is ***scientific merit***, not only impact…
  - Merit/quality is not completely correlated with impact



Pitfall #2: Goodhart's/Campell's law

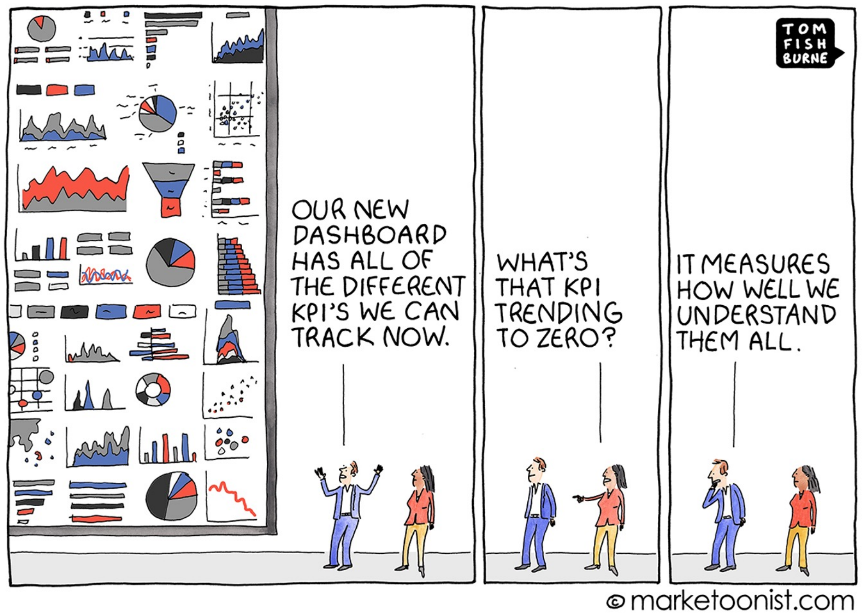
- Scientific impact ***should not be examined through a limited set of measures***.
  - Any individual impact measure has ***limitations***.
  - More measures capture a ***wider range of impact aspects***.
  - **Goodhart's law/Campell's law**: individual measures are vulnerable to attacks & become targets (more measures increased difficulty for attacks)



https://sketchplanations.com/goodharts-law

Pitfall #3: No proper interpretation

- There is ***a multitude of impact measures***.
- In most cases only the measures are provided ***without the proper interpretations***, best practices, etc.
- ***The landscape is confusing*** and often the measures are not properly used.



https://marketoonist.com/2019/11/kpi-overload.html

Part B: Approaches for Estimating the Impact of Papers

Ilias Kanellos (ATHENA RC, Greece)

Background

Wide availability of SKGs

- Large number of scientific papers - publish or perish
- Large number of paper impact assessment methods in literature
  - Many **share similar concepts** and ideas

Background

Wide availability of SKGs

- Large number of scientific papers - publish or perish
- Large number of paper impact assessment methods in literature
    - Many **share similar concepts** and ideas

Different methods evaluated based on

- Different **goals**
- Different **datasets**

Background

Wide availability of SKGs

- Large number of scientific papers - publish or perish
- Large number of paper impact assessment methods in literature
  - Many **share similar concepts** and ideas

Different methods evaluated based on

- Different **goals**
- Different **datasets**

Unclear which method to choose and **under which circumstances**

How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019

Non-Linear PageRank
SPR
SCEAS
Focused PageRank
PrestigeRank
Weighted Citation

How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019

Retained Adjacency Matrix
Timed PageRank
Effective Contagion Matrix
NewRank
NTUWeightedPR
EWPR

Non-Linear PageRank
SPR
SCEAS
Focused PageRank
PrestigeRank
Weighted Citation

THE WEB CONFERENCE ACM

How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019

Non-Linear PageRank
SPR
SCEAS
Focused PageRank
PrestigeRank
hted Citation

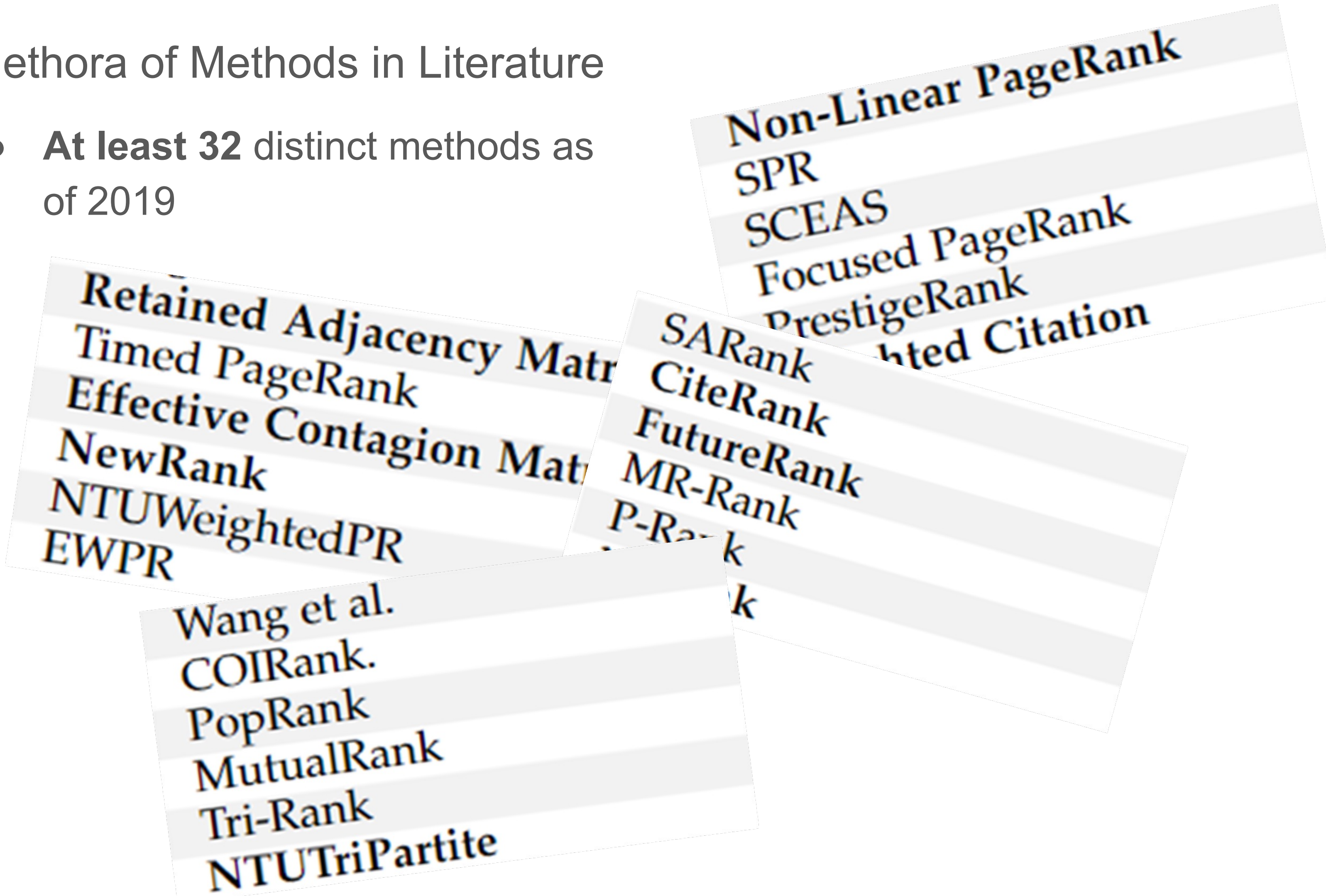
Retained Adjacency Matr
Timed PageRank
Effective Contagion Mat
NewRank
NTUWeightedPR
EWPR

SARank
CiteRank
FutureRank
MR-Rank
P-Rank
YetRank

How to Assess Impact?

Plethora of Methods in Literature

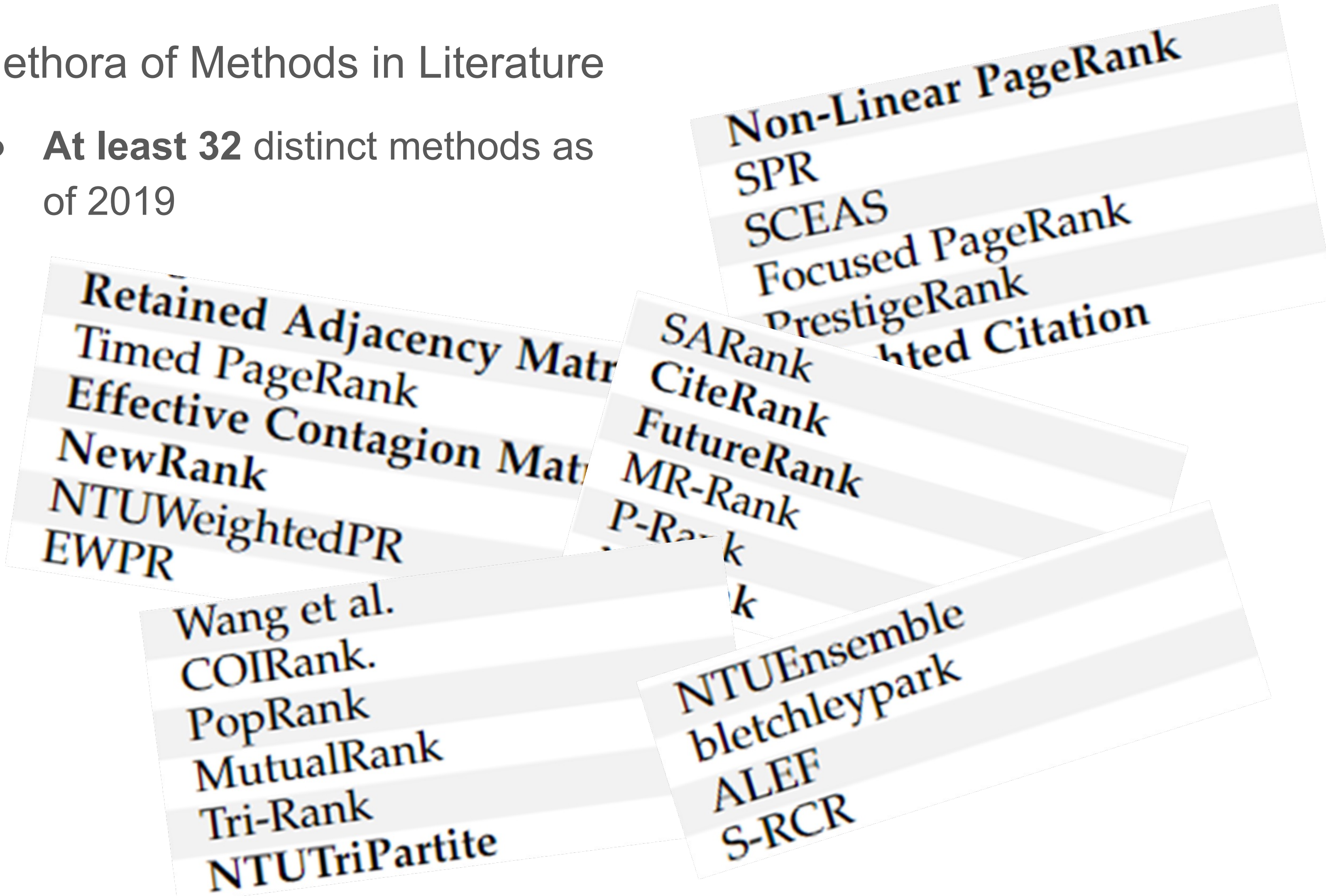
- **At least 32** distinct methods as of 2019



How to Assess Impact?

Plethora of Methods in Literature

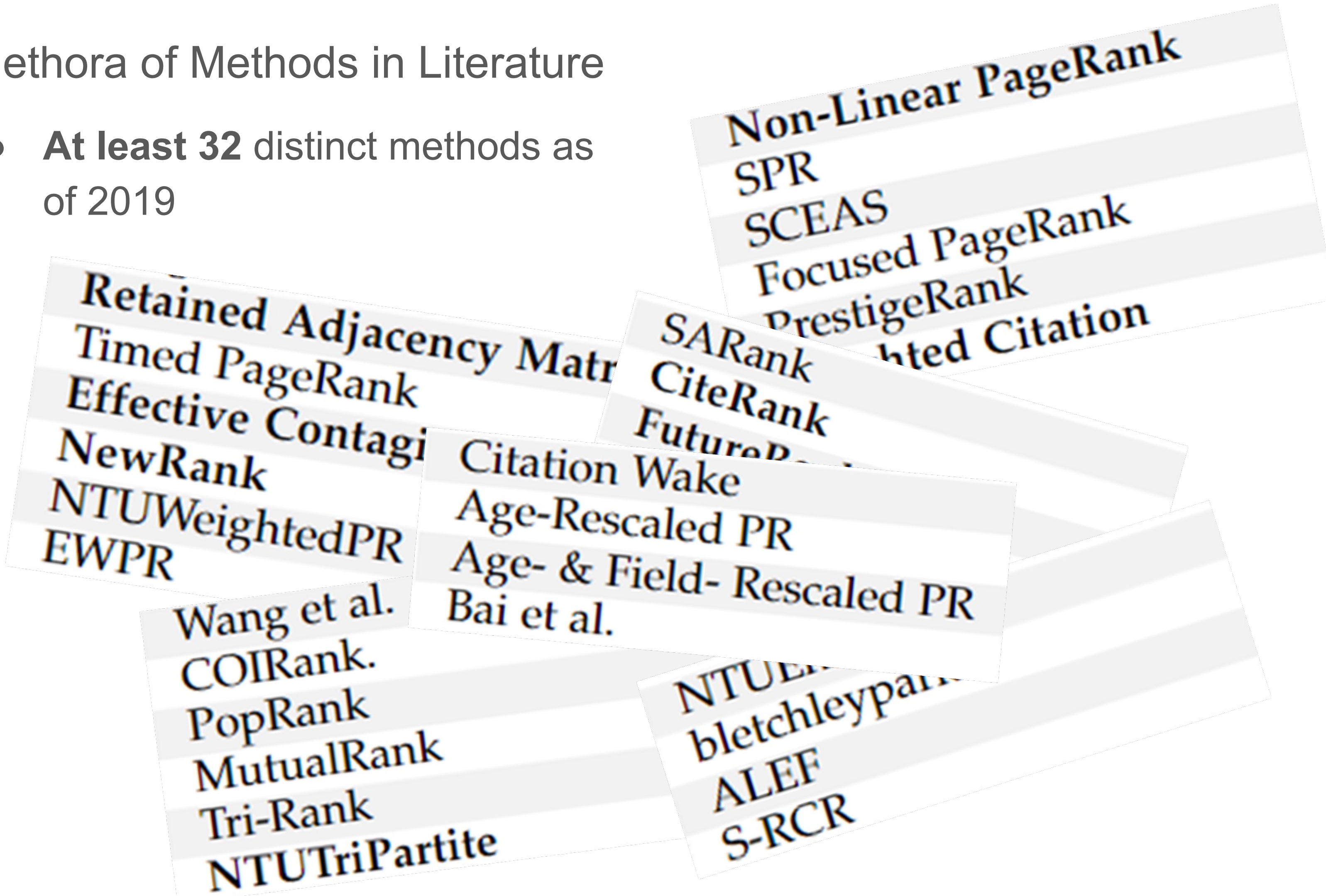
- **At least 32** distinct methods as of 2019



How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019



How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019

Problem dependent

- No clear definition of impact[1]
  - Defined in many different ways

1. Bollen J, Van de Sompel H, Hagberg A, Chute R. A principal component analysis of 39 scientific impact measures. PloS one. 2009 Jun 29;4(6):e6022.

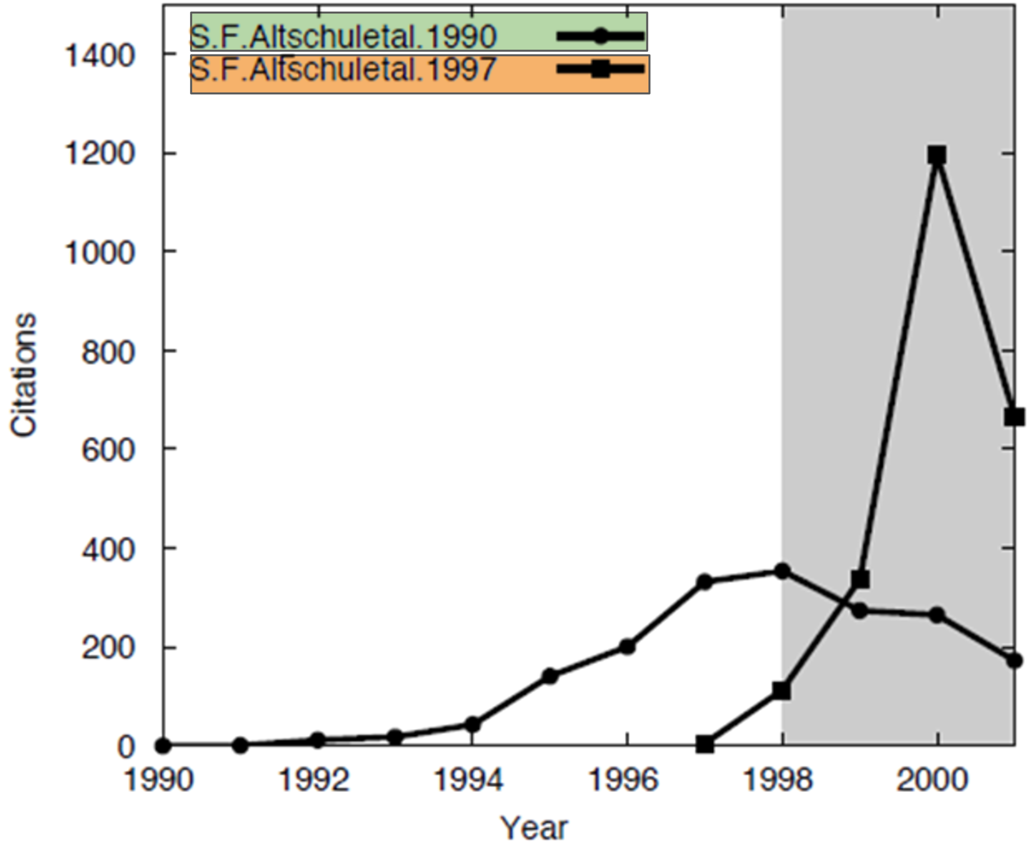
How to Assess Impact?

Plethora of Methods in Literature

- **At least 32** distinct methods as of 2019

Problem dependent

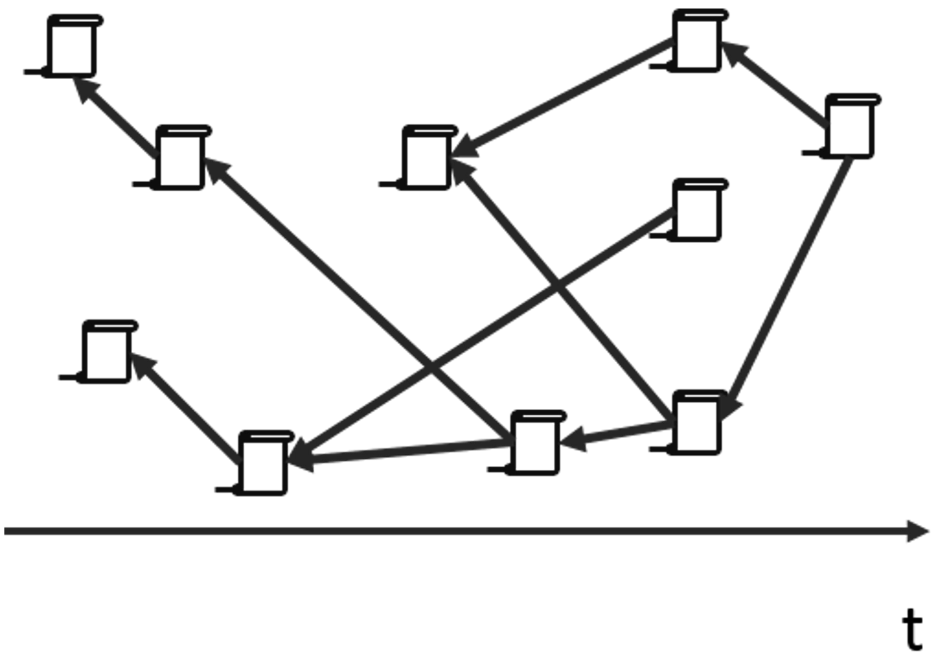
- No clear definition of impact[1]
  - Defined in many different ways
- At least two impact aspects
  - Influence - long term impact
  - Popularity - short term impact



1. Bollen J, Van de Sompel H, Hagberg A, Chute R. A principal component analysis of 39 scientific impact measures. PloS one. 2009 Jun 29;4(6):e6022.

Ranking in Citation Networks

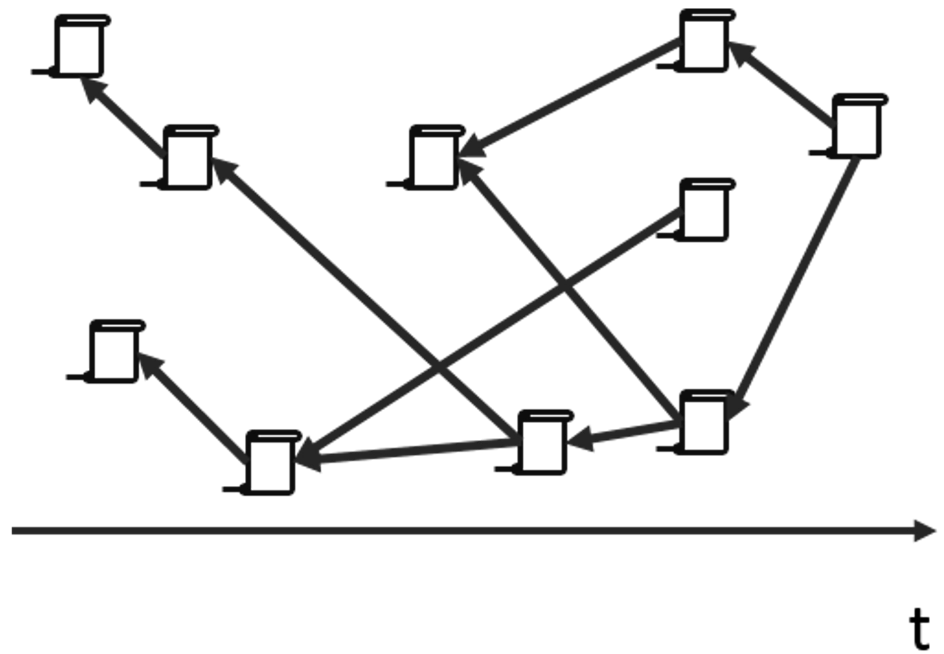
Impact Assessment expressed as **Ranking Problem**



Ranking in Citation Networks

Impact Assessment expressed as **Ranking Problem**

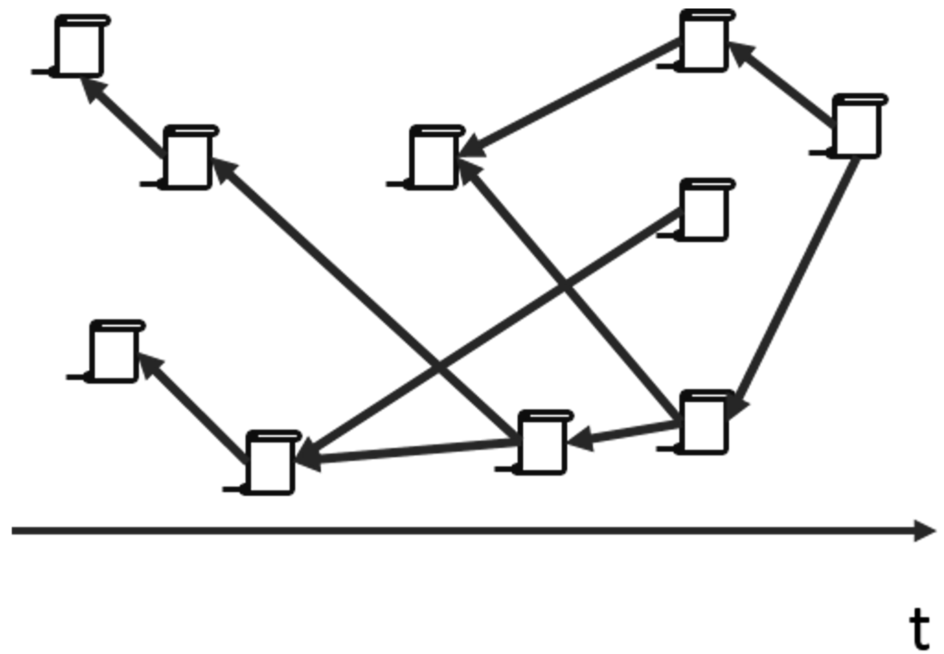
- Impact **assessed comparatively** based on score (e.g., Citation Count)



Ranking in Citation Networks

Impact Assessment expressed as **Ranking Problem**

- Impact **assessed comparatively** based on score (e.g., Citation Count)
- Other **network centrality measures** can be impact proxies
- Much literature **analyzes citation networks** in different ways to assess paper impact



Citation Networks

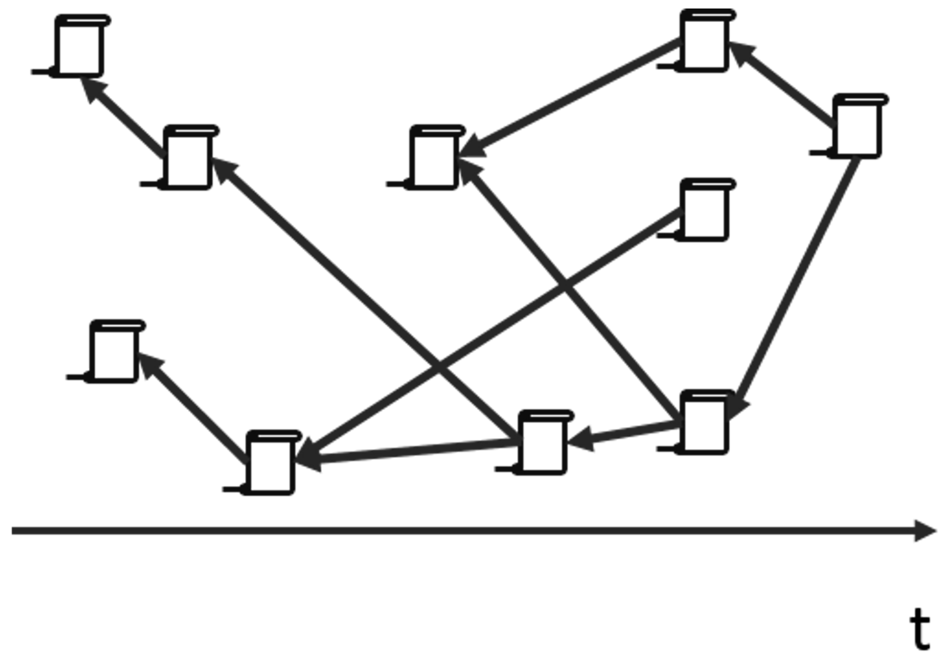
Basic Concepts

Citation Network is a **Graph** with

- Papers as Nodes
- References as edges

References **point backwards** in time

No cycles expected



Citation Networks

Basic Concepts

Citation Network is a **Graph** with

- Papers as Nodes
- References as edges

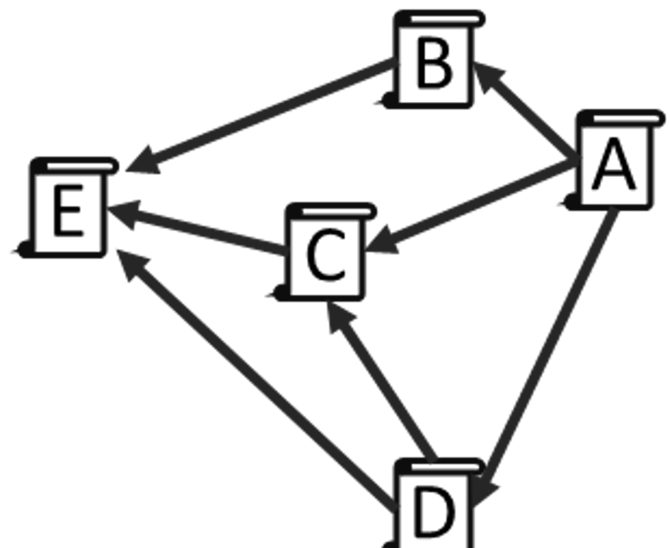
Paper i denoted by p_i

Citations Represented by

Citation Matrix A

$$\boldsymbol{A}: A[i,j] = 1 \; iff \; p_j \rightarrow p_i$$

$$\begin{array}{cccccc} & A & B & C & D & E \\ & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{A} = & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} \end{array}$$



Citation Networks

Basic Concepts

Citation Network is a **Graph** with

- Papers as Nodes
- References as edges

Paper $\boldsymbol{i}$ denoted by p_i

Citations Represented by

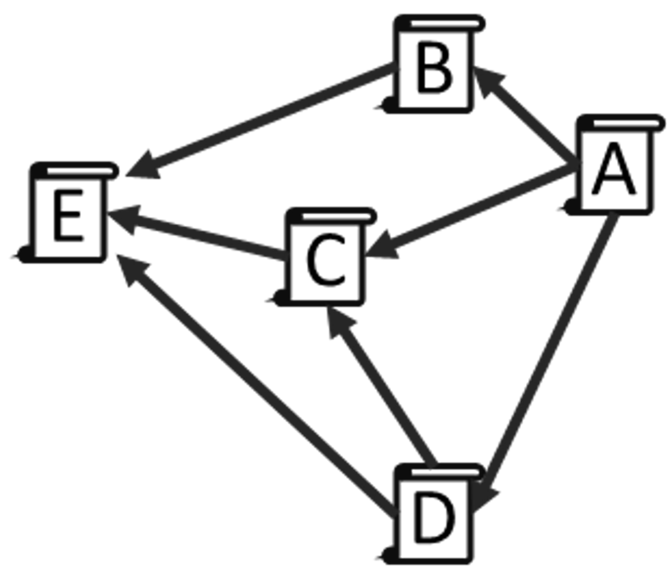
Citation Matrix A

- $\boldsymbol{A}$: $A[i,j] = 1 \ iff \ p_j \rightarrow p_i$

Stochastic Matrix S*

- $\boldsymbol{S}$: $S[i,j] = \dfrac{1}{k} \ iff \ p_j \rightarrow p_i, \ p_j \ cites \ k \ papers$

$$
\begin{array}{cccccc}
 & A & B & C & D & E \\
 & 0 & 0 & 0 & 0 & 1/5 \\
 & 1/3 & 0 & 0 & 0 & 1/5 \\
S = & 1/3 & 0 & 0 & 1/2 & 1/5 \\
 & 1/3 & 0 & 0 & 0 & 1/5 \\
 & 0 & 1 & 1 & 1/2 & 1/5
\end{array}
$$



*Sub-stochastic based on formula. Add 1/N for *dangling nodes*

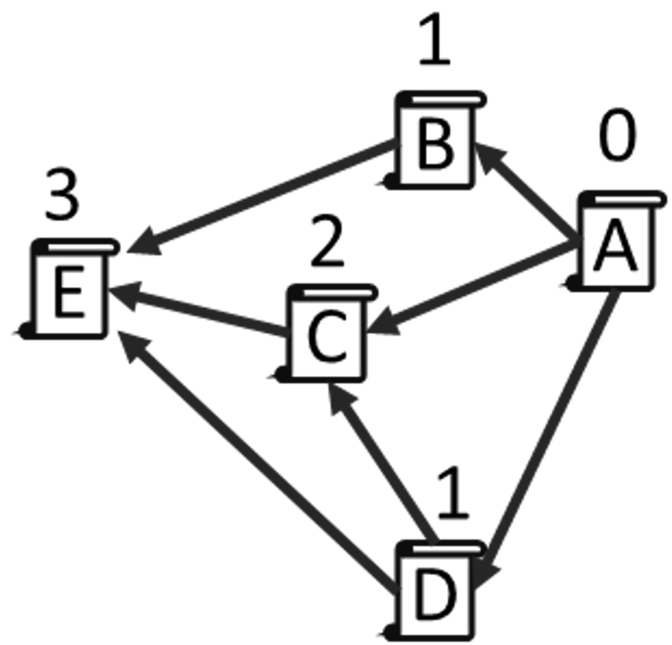
Common Centralities I

Citation Counts

Network centrality measures ~ impact proxies

De facto traditional measure of Scientific Impact

$$
\begin{array}{cccccc}
 & A & B & C & D & E \\
 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
 & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\boldsymbol{A} = & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\
 & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
 & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0}
\end{array}
$$



Common Centralities I

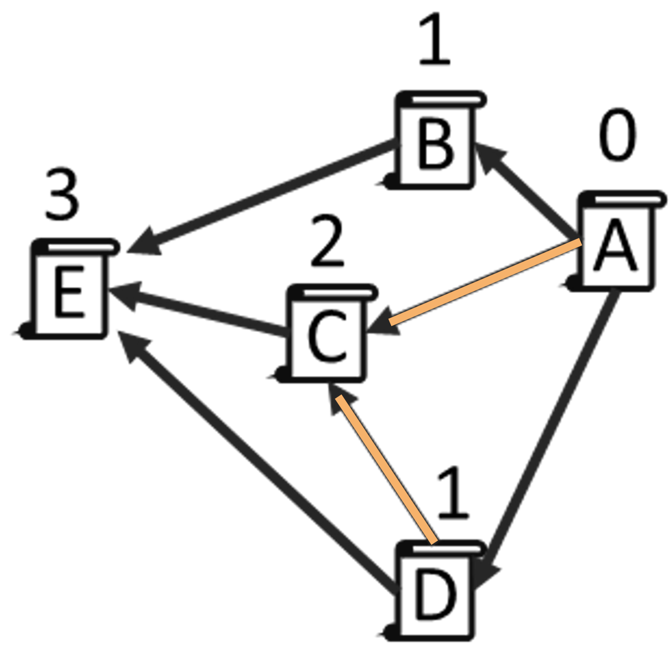
Citation Counts

Network centrality measures ~ impact proxies

De facto traditional measure of Scientific Impact

In terms of Citation Matrix **A**, citation count for paper p_i given as sum over all *j* for row *i*

$$
\begin{array}{cccccc}
 & A & B & C & D & E \\
 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
 & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
A = & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\
 & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
 & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0}
\end{array}
$$



Common Centralities II

PageRank

"A high impact paper is **cited by other** high impact papers"

- Distinguish citing papers by their impact

Common Centralities II

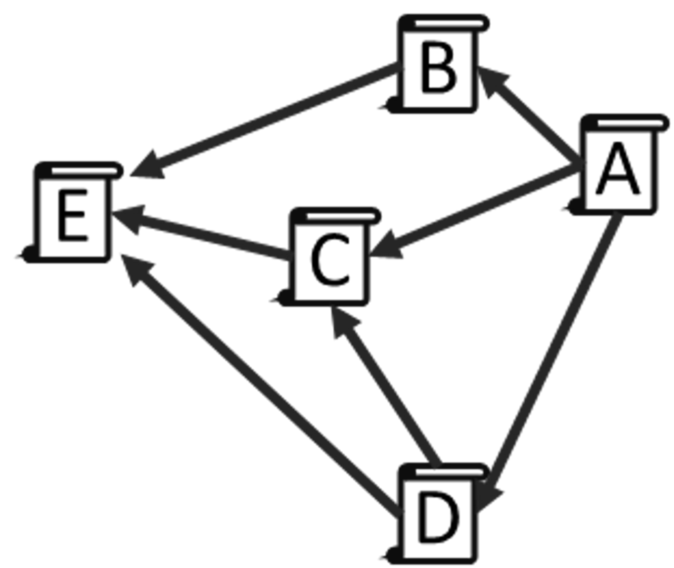
PageRank

“A high impact paper is **cited by other** high impact papers”

- Distinguish citing papers by their impact

$$\begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ S = \begin{bmatrix} 0 & 0 & 0 & 0 & 1/5 \\ 1/3 & 0 & 0 & 0 & 1/5 \\ 1/3 & 0 & 0 & 1/2 & 1/5 \\ 1/3 & 0 & 0 & 0 & 1/5 \\ 0 & 1 & 1 & 1/2 & 1/5 \end{bmatrix} \end{array}$$

$$PR(p_i) = a \sum_j S[i,j] PR(p_j) + (1-a)\frac{1}{N}$$



Common Centralities II

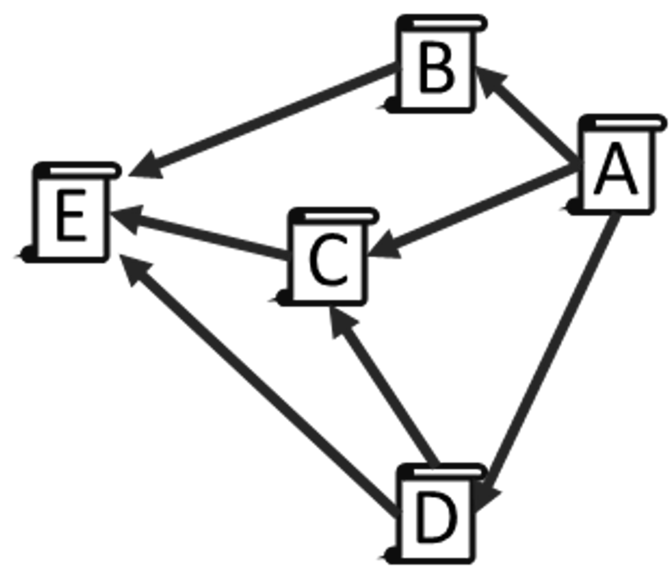
PageRank

"A high impact paper is **cited by other** high impact papers"

- Distinguish citing papers by their impact

$$
\begin{array}{cccccc}
 & A & B & C & D & E \\
S = & \left[\begin{array}{ccccc}
0 & 0 & 0 & 0 & 1/5 \\
1/3 & 0 & 0 & 0 & 1/5 \\
1/3 & 0 & 0 & 1/2 & 1/5 \\
1/3 & 0 & 0 & 0 & 1/5 \\
0 & 1 & 1 & 1/2 & 1/5
\end{array}\right. & & & &
\end{array}
$$

$$PR(C) = a\left[\frac{PR(A)}{3} + \frac{PR(D)}{2} + \frac{PR(E)}{5}\right] + (1-a)\frac{1}{5}$$



Common Centralities II

PageRank

“A high impact paper is **cited by other** high impact papers”

- Distinguish citing papers by their impact
- “Random surfer” (researcher) model

$$PR(p_i) = a \sum_j S[i,j] PR(p_j) + (1-a)\frac{1}{N}$$

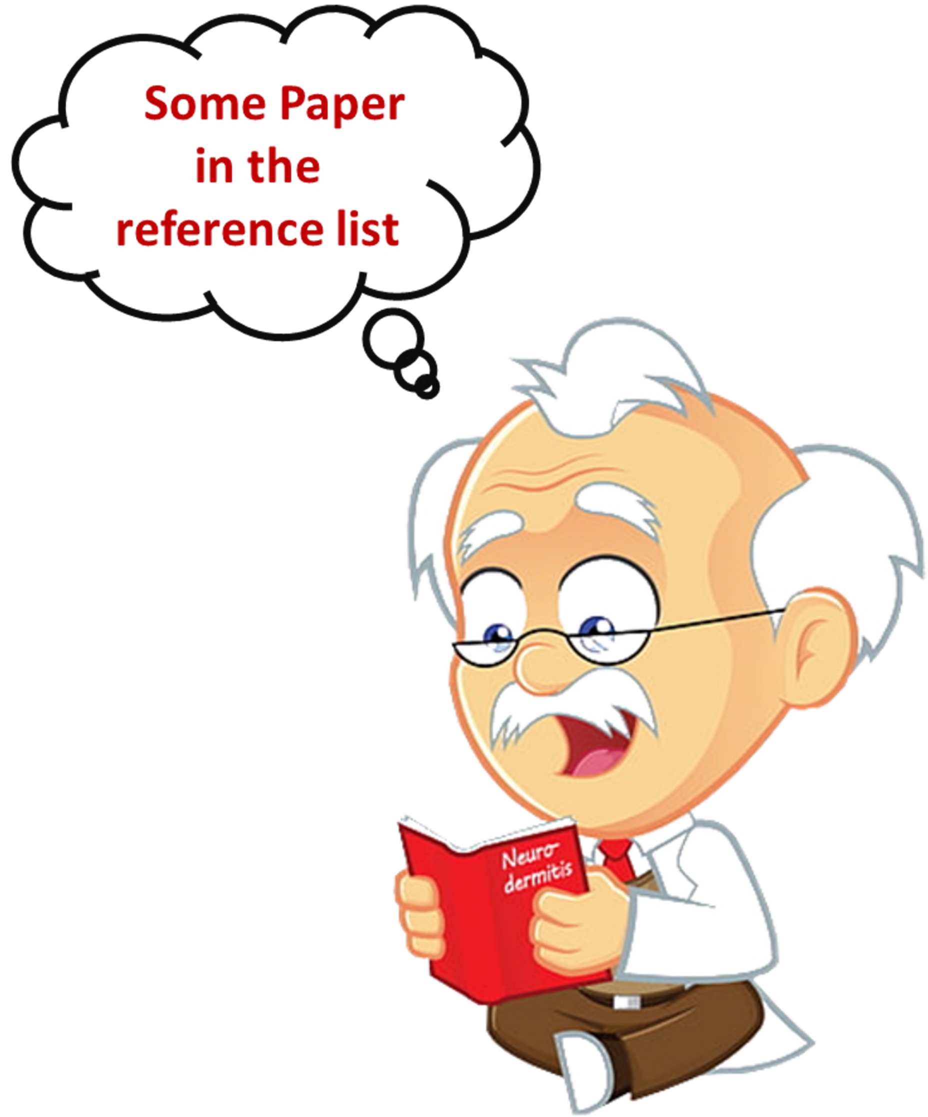
Common Centralities II

PageRank

"A high impact paper is **cited by other** high impact papers"

- Distinguish citing papers by their impact
- "Random surfer" (researcher) model

$$PR(p_i) = \boxed{a \sum_j S[i,j] PR(p_j)} + (1-a)\frac{1}{N}$$



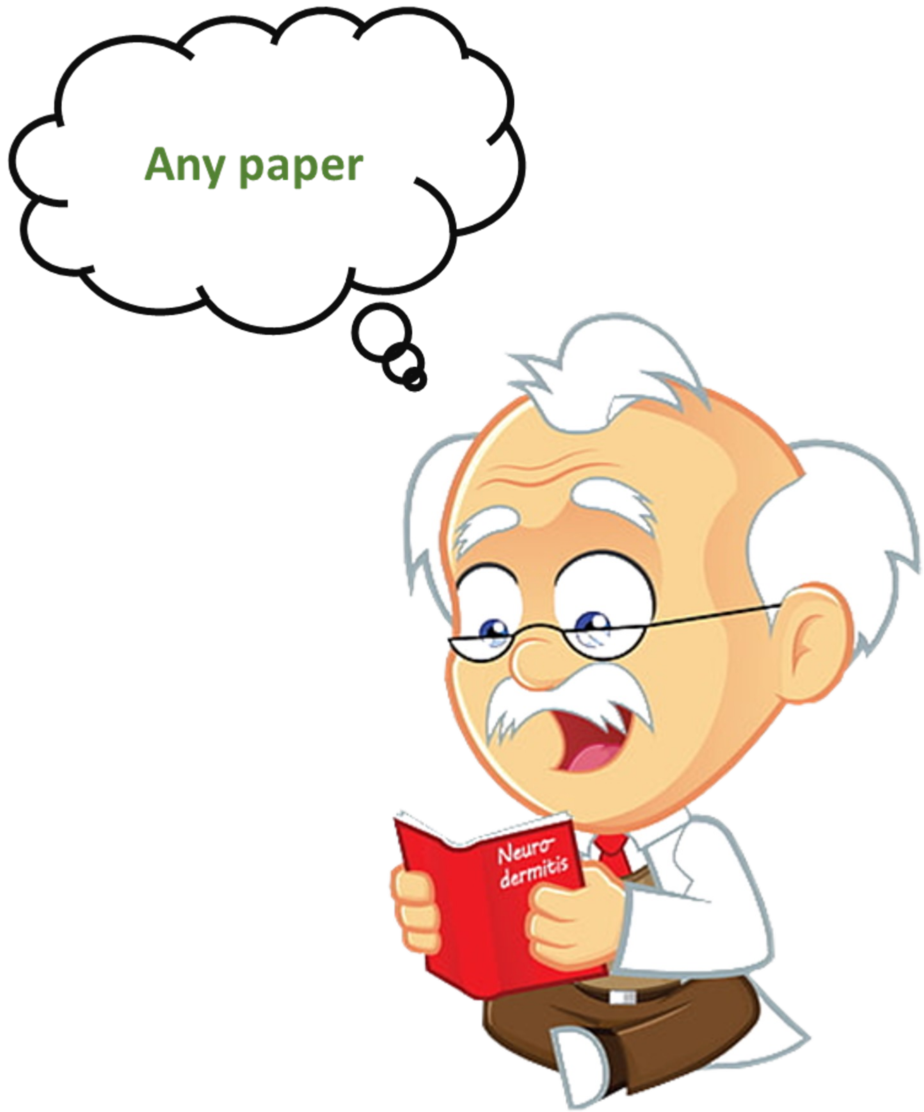
Common Centralities II

PageRank

“A high impact paper is **cited by other** high impact papers”

- Distinguish citing papers by their impact
- “Random surfer” (researcher) model

$$PR(p_i) = a \sum_j S[i,j] PR(p_j) + \boxed{(1-a)\frac{1}{N}}$$



Common Centralities II

PageRank

“A high impact paper is **cited by other** high impact papers”

- Distinguish citing papers by their impact
- “Random surfer” (researcher) model

$$PR(p_i) = a\sum_j S[i,j]PR(p_j) + (1-a)\frac{1}{N}$$

- Early applications on citation networks by Chen et al[1] & Ma et al[2]

1. Chen P, Xie H, Maslov S, Redner S. Finding scientific gems with Google’s PageRank algorithm. Journal of Informetrics. 2007 Jan 1;1(1):8-15.
2. Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Information Processing & Management. 2008 Mar 1;44(2):800-10.

Impact Assessment

Goals and Approaches

Problems

- Citation Count “too democratic” - no differentiation of origin

Approaches

- Balance citations
- Network analyses (e.g., PageRank)
- Weights (e.g., on venues, authors, etc)

Impact Assessment

Goals and Approaches

Problems

- Citation Count "too democratic" - no differentiation of origin

- Older papers have citation headstart / top-ranked papers skewed in favor of old ones

Approaches

- Balance citations
- Network analyses (e.g., PageRank)
- Weights (e.g., on venues, authors, etc)

- Time-awareness
- Exponential decay functions
- Re-scaling / normalizations

Impact Assessment
Goals and Approaches

Problems

- Citation Count “too democratic” - no differentiation of origin

- Older papers have citation headstart / top-ranked papers skewed in favor of old ones

- Avoid “malicious manipulations” and/or “noise”

Approaches

- Balance citations
- Network analyses (e.g., PageRank)
- Weights (e.g., on venues, authors, etc)

- Time-awareness
- Exponential decay functions
- Re-scaling / normalizations

- Neglect self citations
- Consider citing-cited paper similarities

Impact Assessment

Goals and Approaches

Problems

- Citation Count "too democratic" - no differentiation of origin
- Older papers have citation headstart / top-ranked papers skewed in favor of old ones
- Avoid "malicious manipulations" and/or "noise"

Approaches

- Balance citations
- Network analyses (e.g., PageRank)
- Weights (e.g., on venues, authors, etc)

- Time-awareness
- Exponential decay functions
- Re-scaling / normalizations

- Neglect self citations
- Consider citing-cited paper similarities

?

And others…

Classification of Methods in Literature[1]

Method	Basic PR variants	Time Aware		Metadata		Multiple Networks	Ensemble	Other
		Network Matrix	Landing Probability	Venue	Author			
Non-Linear PageRank	✓							
SPR	✓							
SCEAS	✓							
Focused PageRank	✓							
PrestigeRank	✓							
Weighted Citation		✓		✓				
Retained Adjacency Matrix		✓						
Timed PageRank		✓		✓	✓			
Effective Contagion Matrix		✓						
NewRank		✓	✓					
NTUWeightedPR		✓	✓	✓	✓			
EWPR		✓		✓	✓		✓	
SARank		✓		✓	✓		✓	
CiteRank			✓					
FutureRank			✓		✓	✓		
MR-Rank		✓		✓		✓		
P-Rank				✓	✓	✓		
YetRank			✓	✓				
Wang et al.			✓	✓	✓	✓		
COIRank.			✓	✓	✓	✓		
PopRank						✓		
MutualRank						✓		
Tri-Rank				✓	✓	✓		
NTUTriPartite				✓	✓	✓	✓	
NTUEnsemble		✓	✓	✓	✓	✓	✓	
bletchleypark		✓		✓	✓		✓	
ALEF					✓		✓	
S-RCR								✓
Citation Wake								✓
Age-Rescaled PR								✓
Age- & Field- Rescaled PR								✓
Bai et al.								✓

1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-based ranking of scientific publications: a survey and experimental evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13;33(4):1567-84.

Classification I

Data leveraged

Citations only

- Citation Count, PageRank

Paper Metadata

- Publication Venues and/or Author Information
  - Others options (e..g, institution-based info)

Publication time-based metadata (weights)

- Paper age
  - When was a paper **published**
- Citation age
  - When was a paper **cited**
- Citation gap
  - How much **time passed** when a paper was cited since its publication

Classification II

Computational Model

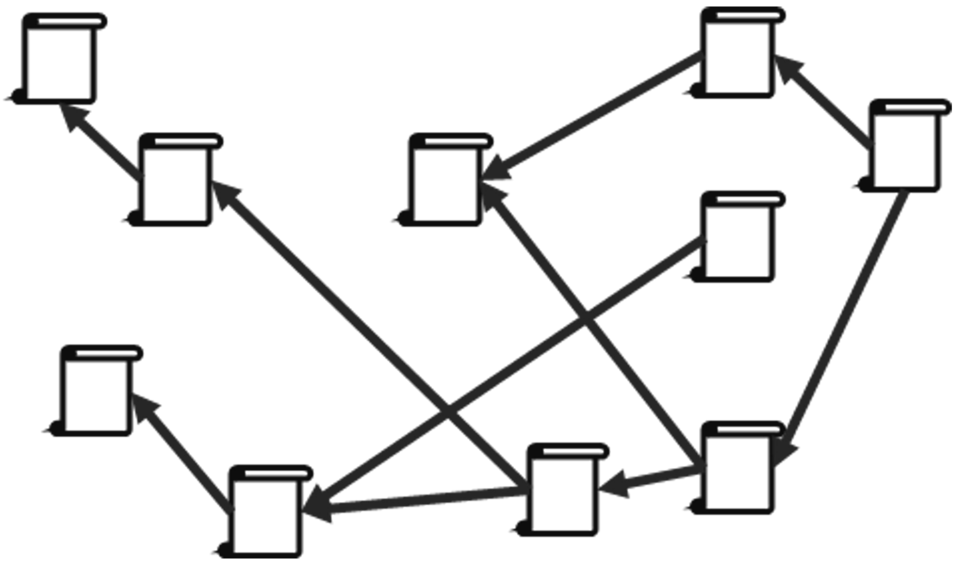
Citation Count

PageRank

Heterogeneous Networks

Ensemble Methods

Other Approaches

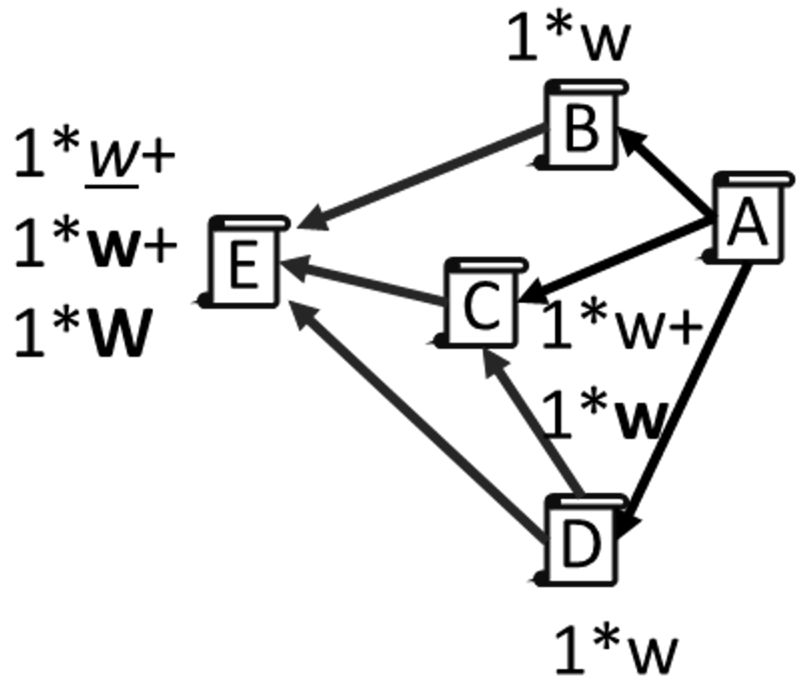


Classification II

Computational Model

Citation Count

- Use **only** (direct) **citations**
- Or **apply weights** on citations (e.g., based on publication venues, based on authors, etc)



PageRank

Heterogeneous Networks

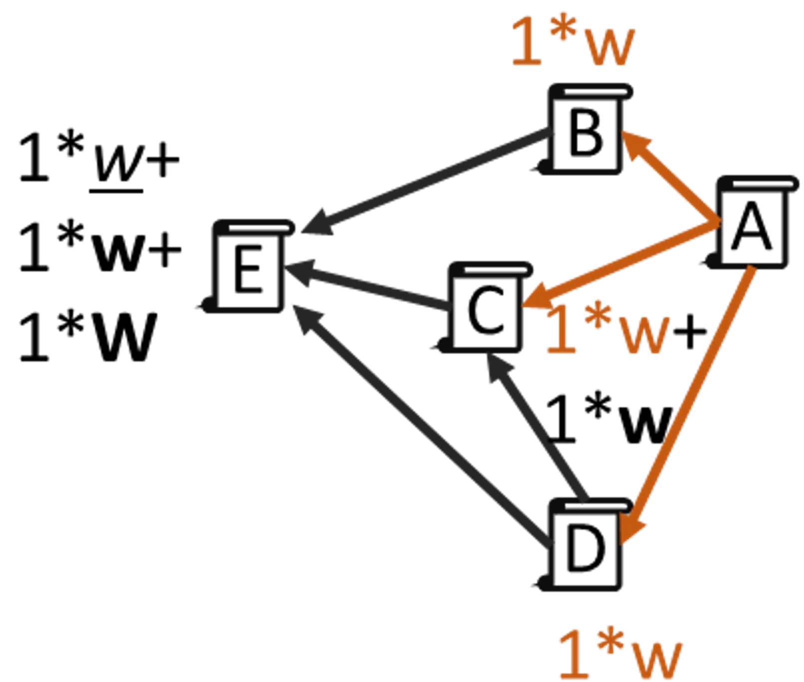
Ensemble Methods

Other Approaches

Classification II
Computational Model

Citation Count

- Use **only** (direct) **citations**
- Or **apply weights** on citations (e.g., based on publication venues, based on authors, etc)
- E.g., citations from A have weight w



PageRank

Heterogeneous Networks

Ensemble Methods

Other Approaches

Citation Count-based Approaches

Example Methods

Weighted Citation[1]

- Weigh citations based on **journal prestige**
    - Weights by **Article Influence Score**, function of Eigenfactor
    - EF: Eigenfactor of A’s Journal is PR-like score on **journal networks**
    - a: fraction of articles in J over a time year window

$$AI_A = 0.01 \frac{EF_{JA}}{a_{JA}}$$

1. Yan E, Ding Y. Weighted citation: An indicator of an article's prestige. Journal of the American Society for Information Science and Technology. 2010 Aug;61(8):1635-43.

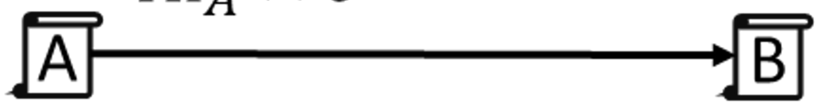
Citation Count-based Approaches

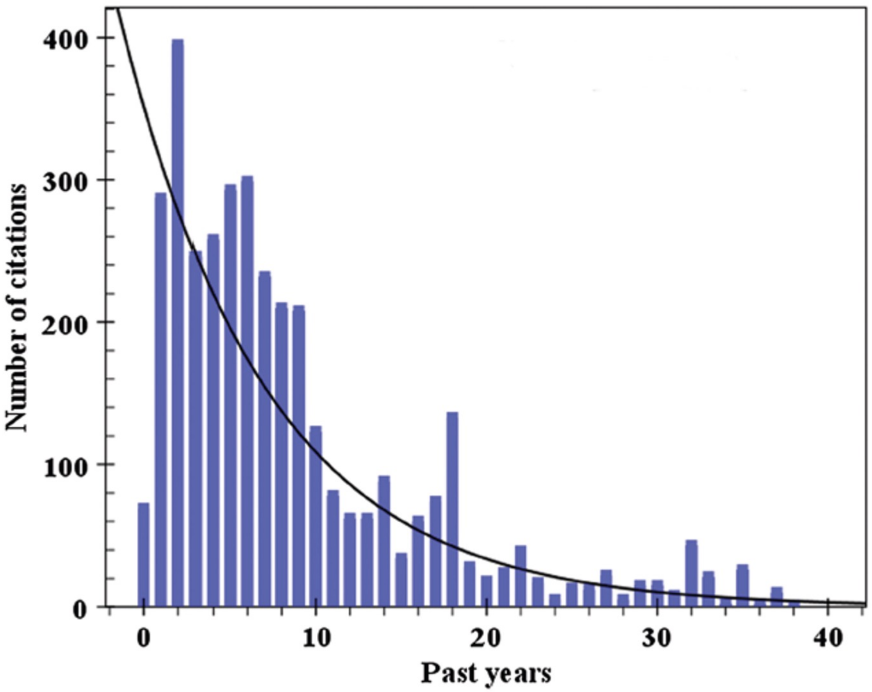
Example Methods

Weighted Citation[1]

- Weigh citations based on **journal prestige**
- Weigh citations based on **"quickness"** (**citation gap**)
  - "Quick citations" considered to convey
    - Important breakthroughs
    - Authority authors
  - f(x)~e−0.117x
    - Based on empirical citation data

$$AI_A \times e^{-0.117(t_A - t_B)}$$





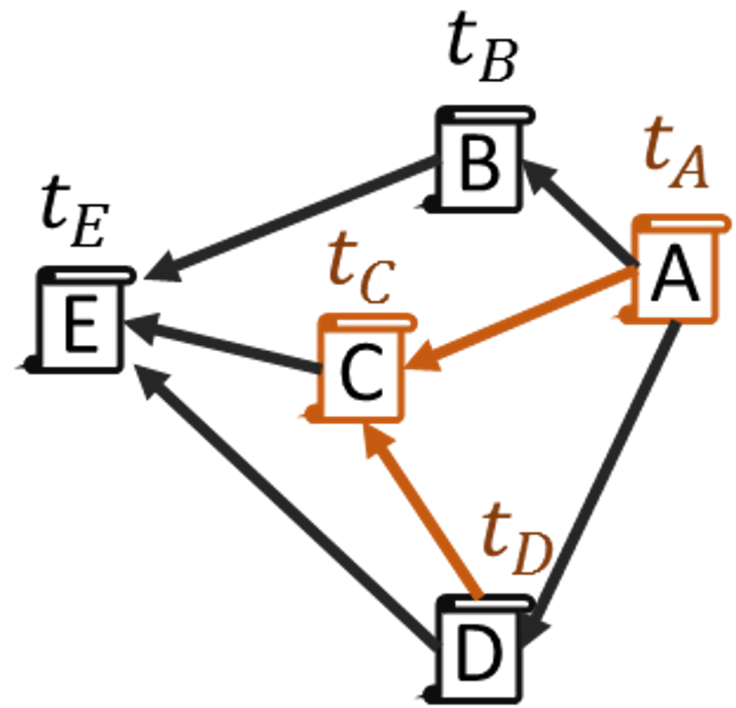
1. Yan E, Ding Y. Weighted citation: An indicator of an article's prestige. Journal of the American Society for Information Science and Technology. 2010 Aug;61(8):1635-43.

Citation Count-based Approaches

Example Methods

Weighted Citation[1]

- Weigh citations based on **journal prestige**
- Weigh citations based on **"quickness"** (**citation gap**)



Example

- Due to nature of citation network
  $t_A > t_C, \quad t_D > t_C$
- Longer citation gaps decrease weight

$$WeightedCitation(C) = AI_A \times e^{-0.117(t_A - t_C)} + AI_D \times e^{-0.117(t_D - t_C)}$$

1. Yan E, Ding Y. Weighted citation: An indicator of an article's prestige. Journal of the American Society for Information Science and Technology. 2010 Aug;61(8):1635-43.

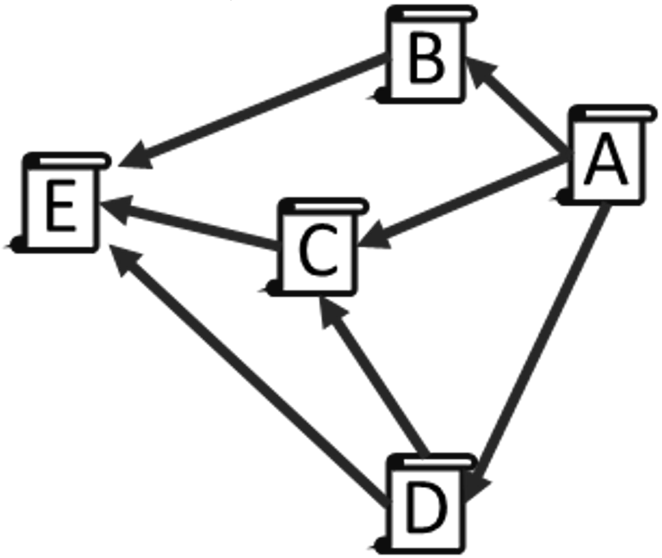
Citation Count-based Approaches

Example Methods

RAM[1]

- **Recent citations** more important
- Adj. Matrix => Retained Adjacency Matrix **(RAM)**
- $R[i,j] = \gamma^{t_N - t_j}, \gamma \in [0,1]$
- N = current year

$$A = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{matrix}$$



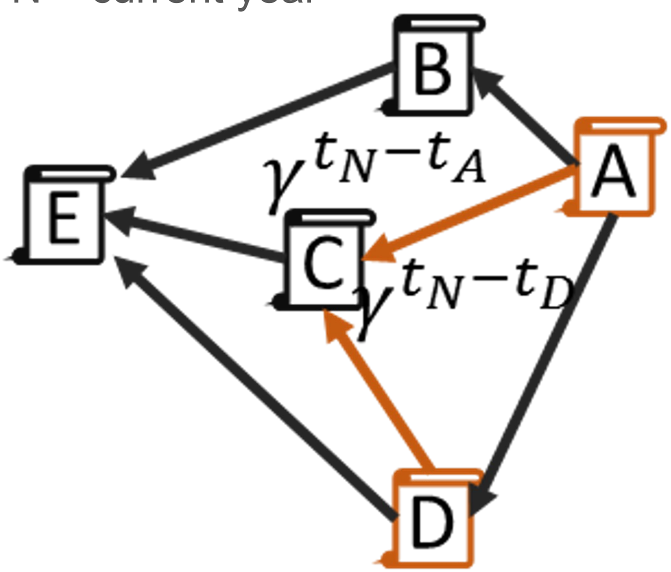
1. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.

Citation Count-based Approaches

Example Methods

RAM[1]

- **Recent citations** more important
- Adj. Matrix => Retained Adjacency Matrix **(RAM)**
- $R[i,j] = \gamma^{t_N - t_j}, \gamma \in [0,1]$
- N = current year



$$A = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{matrix}$$

$$R = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ \gamma^{t_N - t_A} & 0 & 0 & 0 & 0 \\ \gamma^{t_N - t_A} & 0 & 0 & \gamma^{t_N - t_D} & 0 \\ \gamma^{t_N - t_A} & 0 & 0 & 0 & 0 \\ 0 & \gamma^{t_N - t_B} & \gamma^{t_N - t_C} & \gamma^{t_N - t_D} & 0 \end{matrix}$$

$$RAM(C) = \gamma^{t_N - t_A} + \gamma^{t_N - t_D}$$

1. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.

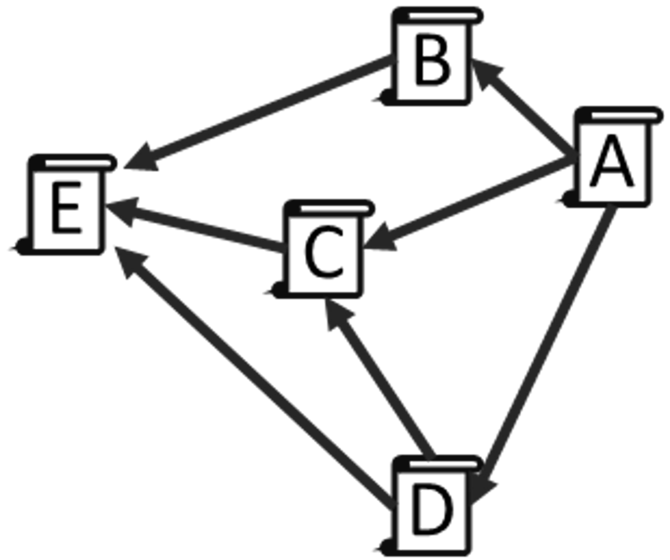
Citation Count-based Approaches

Example Methods (Maybe lose this slide)

ECM[1]

- Expand RAM to calculate **chains of citations**
- **Attenuate** with length

$$ECM[i,j] = \sum_{i=1}^{N-1} a^i R^i, a \in [0,1]$$



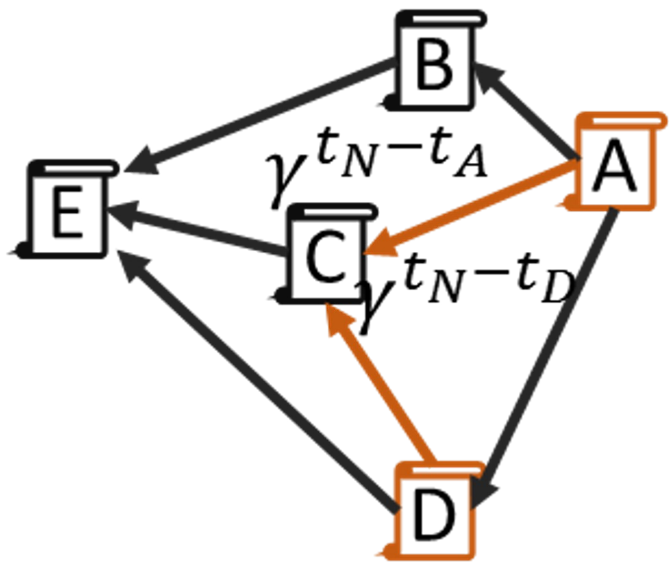
1. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.

Citation Count-based Approaches

Example Methods (Maybe lose this slide)

ECM[1]

- Expand RAM to calculate **chains of citations**
- **Attenuate** with length
- $ECM[i,j] = \sum_{i=1}^{N-1} a^i R^i, a \in [0,1]$



Example

- One-hop paths

$$ECM(C) = \boxed{a\gamma^{t_N - t_A} + a\gamma^{t_N - t_D}} + \cdots$$

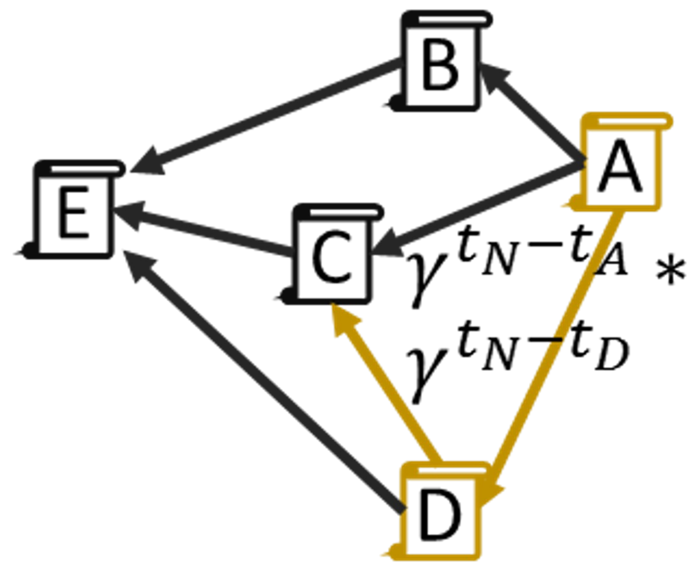
1. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.

Citation Count-based Approaches

Example Methods (Maybe lose this slide)

ECM[1]

- Expand RAM to calculate **chains of citations**
- **Attenuate** with length
- $ECM[i,j] = \sum_{i=1}^{N-1} a^i R^i, a \in [0,1]$



Example

- One-hop paths
- Two-hop paths

$$ECM(C) = \boxed{a\gamma^{t_N - t_A} + a\gamma^{t_N - t_D}} + \boxed{a^2\gamma^{t_N - t_A}\gamma^{t_N - t_D}}$$

1. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.

Classification II

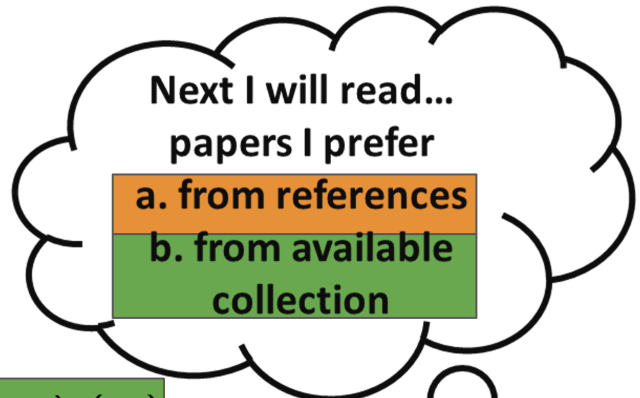
Computational Model

Citation Count

PageRank

- **Modify** random surfer model

$$PR'(p_i) = \boxed{a\sum_j S'[i,j]PR'(p_j)} + \boxed{(1-a)f(p_i)}$$



Heterogeneous Networks

Ensemble Methods

Other Approaches

PageRank
Semantics

PageRank **simulates** “random researcher”

- When reading a particular paper p_i choose
  - With probability **a** another paper **in its reference list**
  - With probability **1-a any paper** in the citation network
- Next paper p_i **depends only on paper** p_j

This behaviour can be modeled by a Finite State **Discrete Markov Chain**

- Transition matrix $G = aS + \frac{(1-a)}{N}J$
  J: matrix of all 1s
- PageRank scores are values of stationary distribution of **G**
- Calculate using **power iteration**

$$\overrightarrow{PR} = G\overrightarrow{PR}$$

$$\overrightarrow{PR_{k+1}} = G\overrightarrow{PR_k}$$

PageRank
Convergence

PageRank vector results from application of power method on **G** matrix

Convergence guaranteed by Perron-Frobenius Theorem[1] when

- Matrix is stochastic (valid by definition for **G**)
- Matrix is irreducible
  - Guaranteed when all states can transition to all other states (all papers "cite" all other papers)
  - Guaranteed for **G**, because all cells > 0, least value $(1-\alpha)/N$
- Matrix is aperiodic
  - Guaranteed by self-loops (i.e., non zero diagonal entries of matrix **G**)
  - Guaranteed by PageRank's random jump vector

1. Langville AN, Meyer CD. Google's PageRank and beyond. Princeton university press; 2011 Jul 1.

PageRank

Convergence Consequences

Define any matrix **S'** which is

- Stochastic
- Instead of 1/k, use **different weights**, as long as matrix stays **column-stochastic**

Add **custom-jump vector** (vanilla PageRank is uniform)

- Ensure **non-zero values in all cells**
  - Choose vector w/ **positive values** on all dimensions
  - Normalize it

Above interventions easily translate to particular "* researcher" behaviour

Any quantity can be **normalized and applied** in to Stochastic Matrix and/or Random jump vector

PageRank

Adjustments to **G** matrix

Focused PageRank[1]

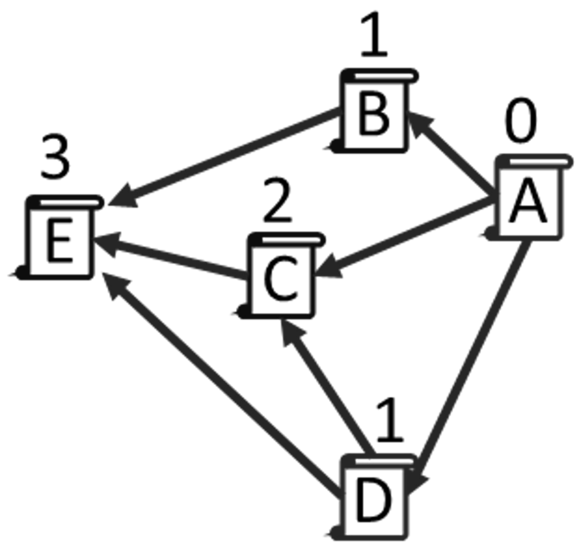
- Balance PR and CC
- Researcher **prefers most cited** among papers in reference list
- Replace 1/k in **S** with

$$\frac{CC(p_i)}{\sum_{i,j \to i} CC(p_j)}$$

$$\begin{array}{cccccc} & A & B & C & D & E \\ S = & \begin{matrix} 0 \\ 1/4 \\ \boxed{2/4} \\ 1/4 \\ 0 \end{matrix} & \begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{matrix} & \begin{matrix} 0 \\ 0 \\ 2/5 \\ 0 \\ 3/5 \end{matrix} & \begin{matrix} 0 \\ 1/7 \\ 2/7 \\ 1/7 \\ 3/7 \end{matrix} \end{array}$$

Example

$$FPR(C) = a\left[\frac{2}{4}FPR(A) + \frac{2}{5}FPR(D) + \frac{2}{7}FPR(E)\right] + \frac{1-a}{5}$$



1. Krapivin M, Marchese M. Focused page rank in scientific papers ranking. InInternational Conference on Asian Digital Libraries 2008 Dec 2 (pp. 144-153). Springer, Berlin, Heidelberg.

PageRank

Adjustments to **G** matrix

Similarity Preferential PageRank[1]

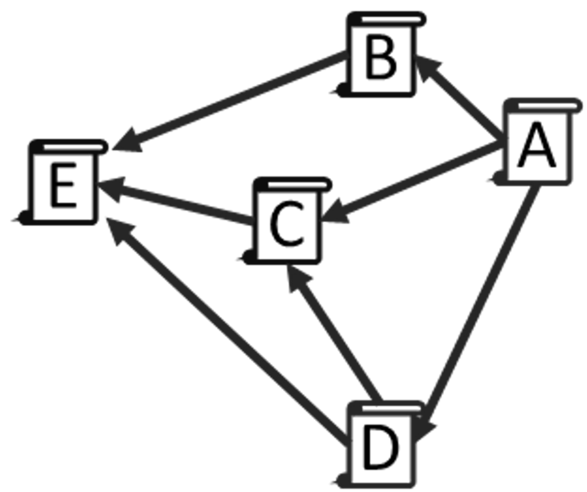
- Avoid malicious manipulations
- Researcher **prefers similar papers**
- Replace 1/k in **S** with

  $$\frac{f_{ij}^{\theta}}{k}, \theta > 0, f_{ij}^{\theta} = \frac{|\tau(i) \cap \tau(j)|}{\sqrt{k_i k_j}}$$

  τ(i) is set of papers cited by pi *

- Similar papers cite similar sets of papers

$$S = \begin{matrix} 0 & 0 & 0 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1/5 \\ 0 & 0 & 0 & \left(1/\sqrt{2}\right)^{\theta}/2 & 1/5 \\ \left(1/\sqrt{5}\right)^{\theta}/3 & 0 & 0 & 0 & 1/5 \\ 0 & 0 & 0 & 0 & 1/5 \end{matrix}$$



Example (no common cited papers among C & A)

$$SPR(C) = a\left[\frac{1}{2\sqrt{2}^{\theta}} SPR(D) + \frac{1}{5} SPR(E)\right] + \frac{1-a}{5}$$

1. Zhou J, Zeng A, Fan Y, Di Z. Ranking scientific publications with similarity-preferential mechanism. Scientometrics. 2016 Feb;106(2):805-16.

* Convergence is shown experimentally - not by Perron - Frobenius theorem

PageRank
Time Aware Approach

CiteRank[1]

- Assumption: researchers start browsing from **recent works**
  - Then follow citations
- **Modify random jump** vector

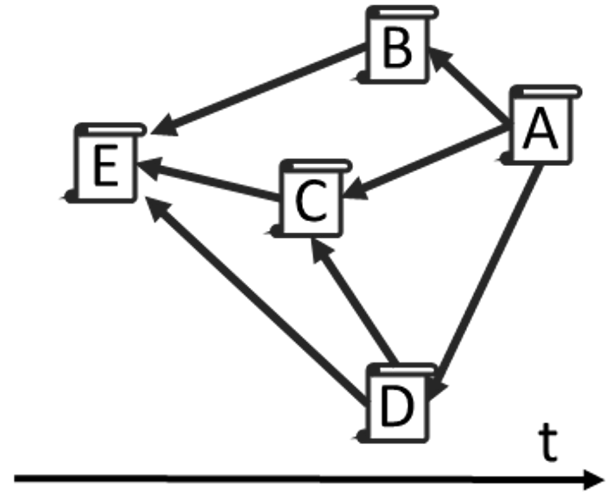
$$\rho_i = e^{\frac{t_i}{\tau_{dir}}}$$

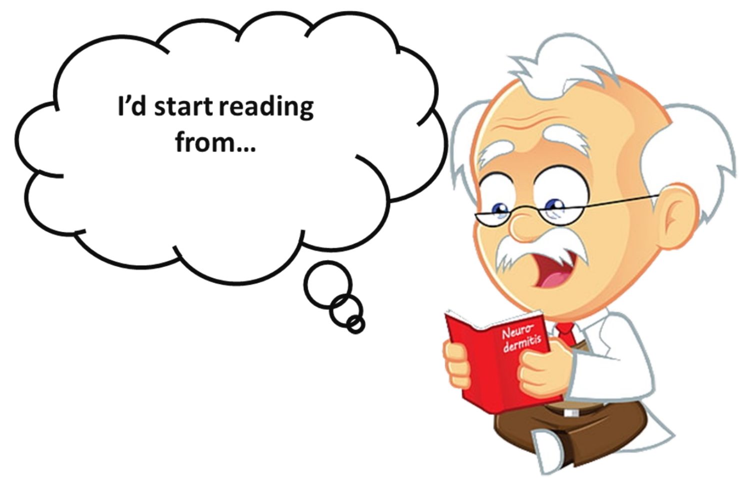
- CiteRank defined as

$$\overrightarrow{CR} = I\vec{\rho} + (1-a)S\vec{\rho} + (1-a)^2 S^2 \vec{\rho} + \ldots +$$

- If $\vec{\rho}$ normalized, rewrite[2] as

$$CR(p_i) = a\sum_j S[i,j]\,CR(p_j) + (1-a)\,\rho_i$$





1. Walker D, Xie H, Yan KK, Maslov S. Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment. 2007 Jun 14;2007(06):P06010.
2. Mariani MS, Medo M, Zhang YC. Identification of milestone papers through time-balanced network centrality. Journal of Informetrics. 2016 Nov 1;10(4):1207-23.

PageRank

Time Aware Approach

CiteRank[1]

- Assumption: researchers start browsing from **recent works**
  - Then follow citations
- **Modify random jump** vector

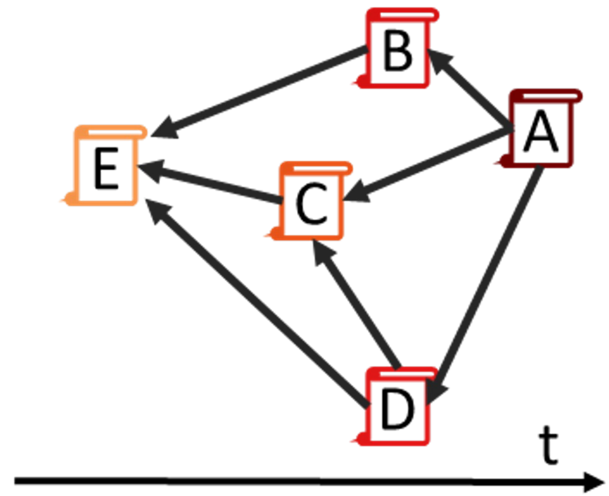
$$\rho_i = e^{\frac{t_i}{\tau_{dir}}}$$

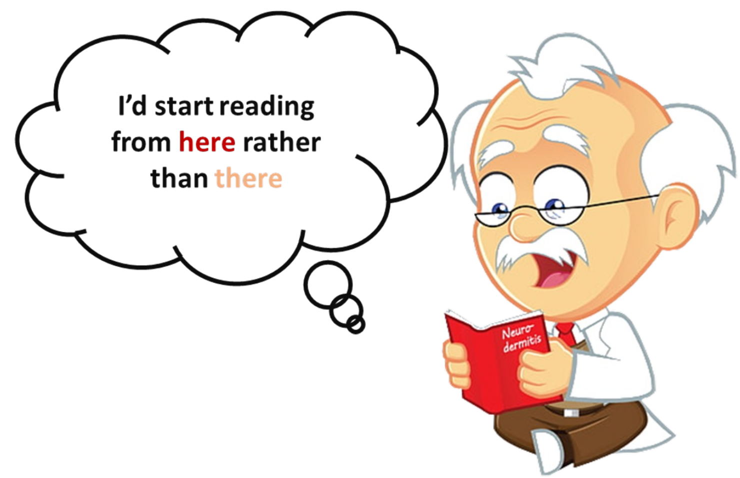
- CiteRank defined as

$$\vec{CR} = I\vec{\rho} + (1-a)S\vec{\rho} + (1-a)^2 S^2 \vec{\rho} + \ldots +$$

- If $\vec{\rho}$ normalized, rewrite[2] as

$$CR(p_i) = a \sum_j S[i,j] CR(p_j) + \boxed{(1-a)\,\rho_i}$$





1. Walker D, Xie H, Yan KK, Maslov S. Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment. 2007 Jun 14;2007(06):P06010.
2. Mariani MS, Medo M, Zhang YC. Identification of milestone papers through time-balanced network centrality. Journal of Informetrics. 2016 Nov 1;10(4):1207-23.

Engineering PageRank

Our time-aware approach

AttRank[1]

- Aim: **current research trends**
- Apply **preferential attachment**
  - Rich get richer
- Intuition: use only **y-most recent years**

$$AR(p_i) = \alpha \sum_j S[i,j] AR(p_j) + \beta \frac{CC_y(p_i)}{\sum_i CC_y(p_i)} + \gamma \frac{e^{-\rho\, t(p_i)}}{\sum_i e^{-\rho\, t(p_i)}}$$

- α+β+γ=1, β & γ normalized
  - **Guarantees** convergence
- Researcher starts reading **recently published**, or **recently cited** papers.

1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Ranking papers by their short-term scientific impact. In2021 IEEE 37th International Conference on Data Engineering (ICDE) 2021 Apr 19 (pp. 1997-2002). IEEE.

Engineering PageRank

Our time-aware approach

AttRank[1]

- Aim: **current research trends**
- Apply **preferential attachment**
  - Rich get richer
- Intuition: use only **y-most recent years**

$$AR(p_i) = \boxed{\alpha \sum_j S[i,j] AR(p_j)} + \beta \frac{CC_y(p_i)}{\sum_i CC_y(p_i)} + \gamma \frac{e^{-\rho\, t(p_i)}}{\sum_i e^{-\rho\, t(p_i)}}$$

- α+β+γ=1, β & γ normalized
  - **Guarantees** convergence
- Researcher starts reading **recently published**, or **recently cited** papers.

1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Ranking papers by their short-term scientific impact. In2021 IEEE 37th International Conference on Data Engineering (ICDE) 2021 Apr 19 (pp. 1997-2002). IEEE.

Engineering PageRank

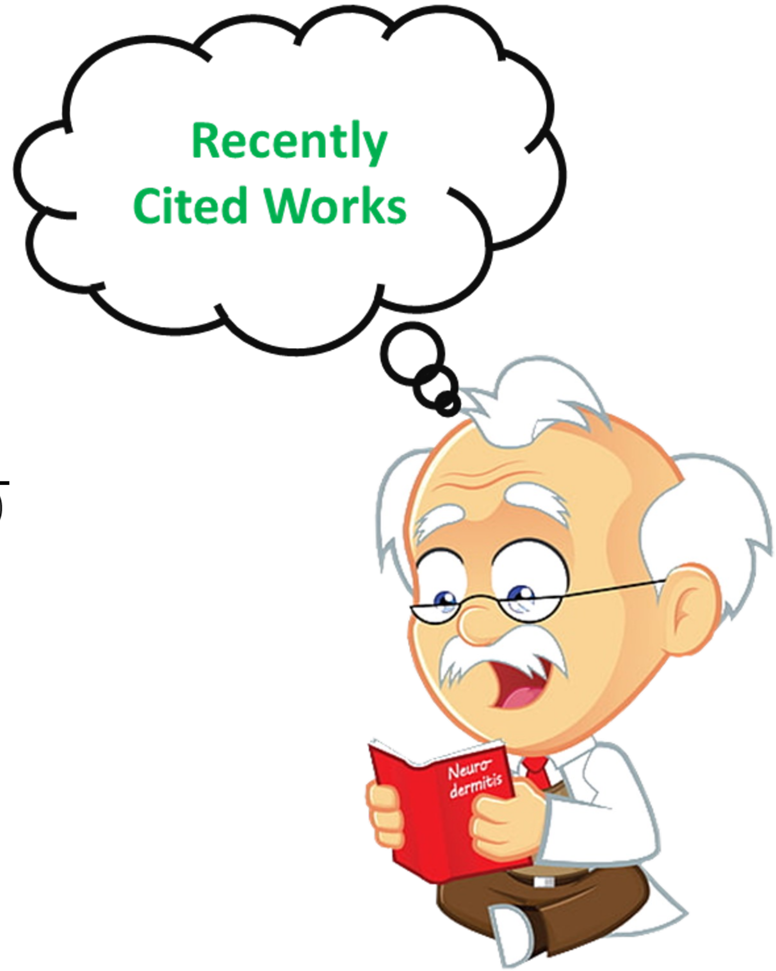
Our time-aware approach

AttRank[1]

- Aim: **current research trends**
- Apply **preferential attachment**
  - Rich get richer
- Intuition: use only **y-most recent years**

$$AR(p_i) = \alpha \sum_j S[i,j] AR(p_j) + \boxed{\beta \frac{CC_y(p_i)}{\sum_i CC_y(p_i)}} + \gamma \frac{e^{-\rho\, t(p_i)}}{\sum_i e^{-\rho\, t(p_i)}}$$

- α+β+γ=1, β & γ normalized
  - **Guarantees** convergence
- Researcher starts reading **recently published,** or **recently cited** papers.



1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Ranking papers by their short-term scientific impact. In2021 IEEE 37th International Conference on Data Engineering (ICDE) 2021 Apr 19 (pp. 1997-2002). IEEE.

Engineering PageRank

Our time-aware approach

AttRank[1]

- Aim: **current research trends**
- Apply **preferential attachment**
  - Rich get richer
- Intuition: use only **y-most recent years**

$$AR(p_i) = \alpha \sum_j S[i,j] AR(p_j) + \beta \frac{CC_y(p_i)}{\sum_i CC_y(p_i)} + \gamma \boxed{\frac{e^{-\rho\, t(p_i)}}{\sum_i e^{-\rho\, t(p_i)}}}$$

- $\alpha+\beta+\gamma=1$, β & γ normalized
  - **Guarantees** convergence
- Researcher starts reading **recently published,** or **recently cited** papers.

1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Ranking papers by their short-term scientific impact. In2021 IEEE 37th International Conference on Data Engineering (ICDE) 2021 Apr 19 (pp. 1997-2002). IEEE.

Classification II

Computational Model

Citation Count

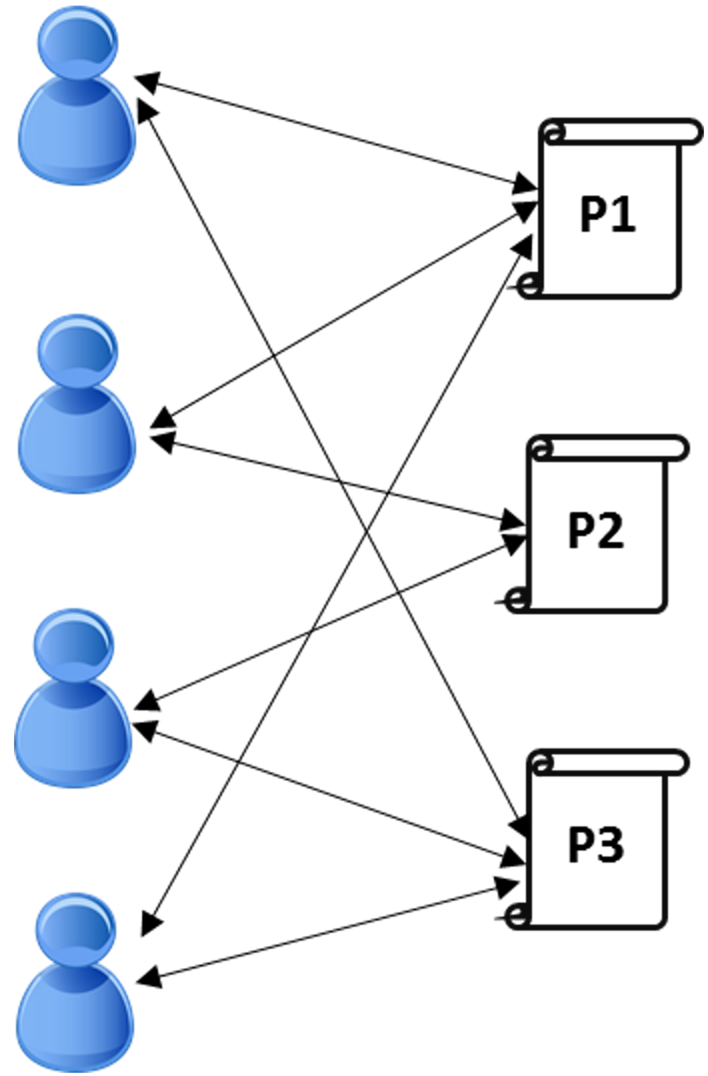
PageRank

Heterogeneous Networks

- Nodes represent **different** types of **entities**
- **Edges** represent **relations** (e.g., paper published in venue)
- Some methods inspired by HITS apply **mutual reinforcement**
- Can provide **rankings of different entities** (e.g., authors and papers)

Ensemble Methods

Other Approaches

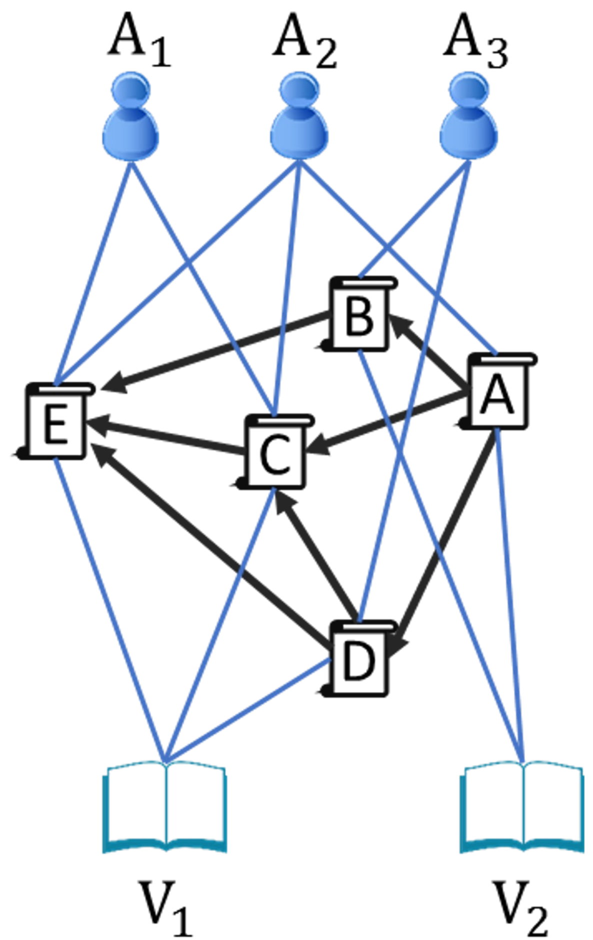


Heterogeneous Networks

Applications

P-Rank[1]

- **Differentiate citations** based on citing papers, journals, authors
- Defines inter- and intra-graph **walks on heterogeneous network**
- **Author scores** based on their papers
- **Venue scores** based on their papers
- **“Random” Jump Vector** based on above, **run PageRank iteration**



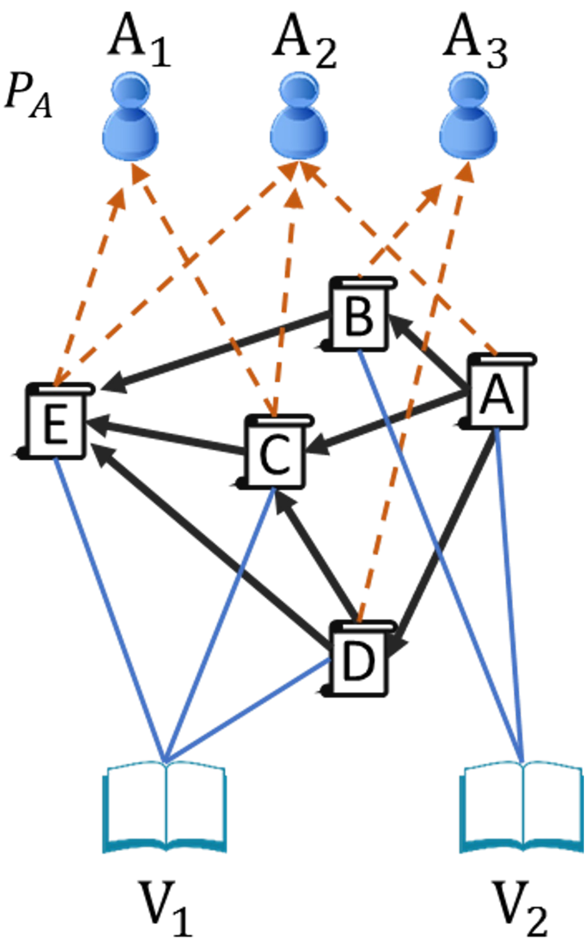
1. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.

Heterogeneous Networks
Applications

P-Rank[1]

- **Differentiate citations** based on citing papers, journals, authors
- Defines inter- and intra-graph **walks on heterogeneous network**
- **Author scores** based on their papers
- **Venue scores** based on their papers
- **"Random" Jump Vector** based on above, **run PageRank iteration**

$$P_{A1} = P_E + P_C$$
$$P_{A2} = P_E + P_C + P_A$$
$$P_{A3} = P_B + P_D$$



1. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.

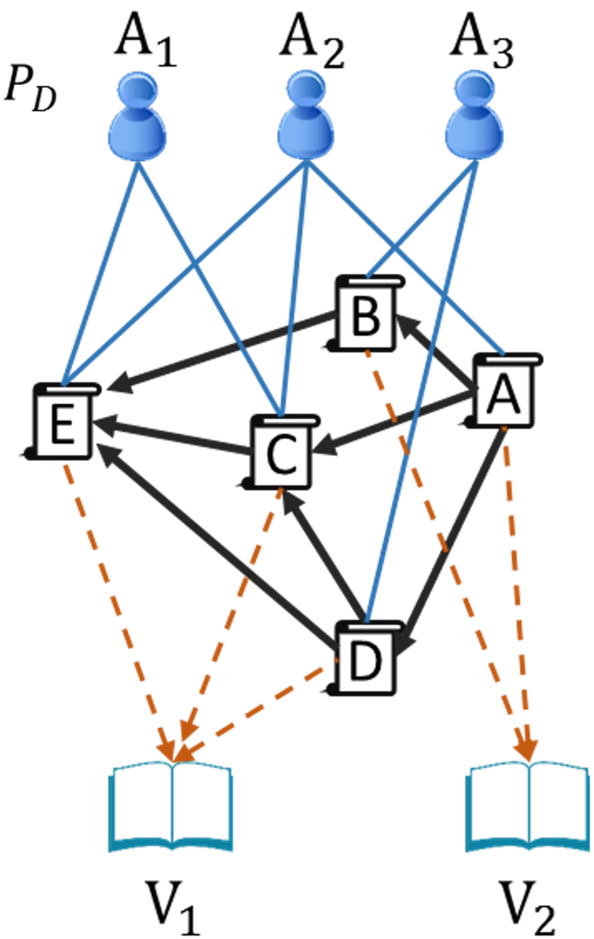
Heterogeneous Networks

Applications

P-Rank[1]

- **Differentiate citations** based on citing papers, journals, authors
- Defines inter- and intra-graph **walks on heterogeneous network**
- **Author scores** based on their papers
- **Venue scores** based on their papers
- **"Random" Jump Vector** based on above, **run PageRank iteration**

$$P_{V1} = P_E + P_C + P_D$$
$$P_{V2} = P_A + P_B$$



1. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.

Heterogeneous Networks

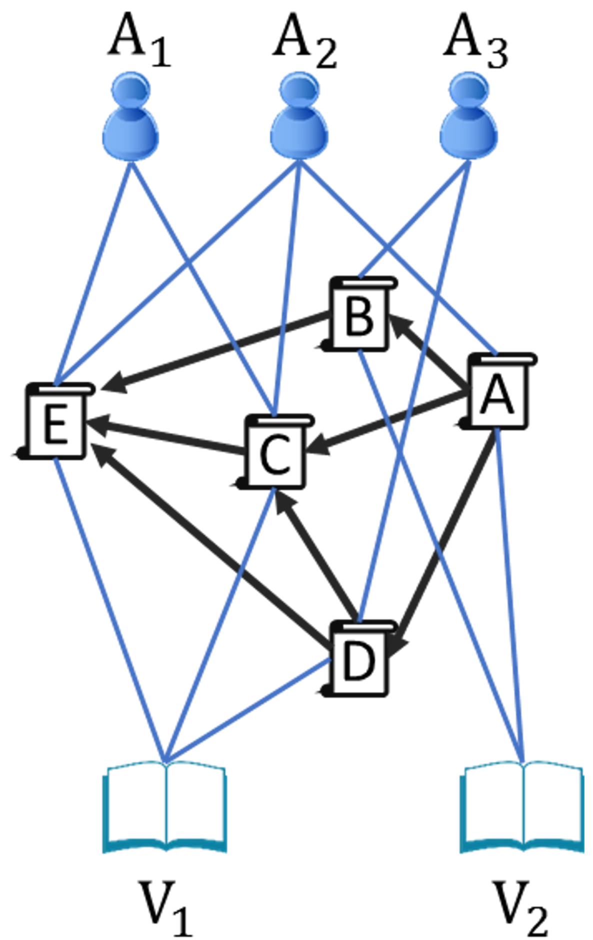
Applications

P-Rank[1]

- **Differentiate citations** based on citing papers, journals, authors
- Defines inter- and intra-graph **walks on heterogeneous network**
- **Author scores** based on their papers
- **Venue scores** based on their papers
- **"Random" Jump Vector** based on above, **run PageRank iteration**

$$P(p_i) = a\sum_j S[i,j]P(p_j) + (1-a)\left[b\sum_{A_i->p_i}\left(\frac{P_{Ai}}{N_{Ai}}\right) + c\sum_{V_i->p_i}\left(\frac{P_{Vi}}{N_{Vi}}\right)\right]$$

$$b + c = 1$$



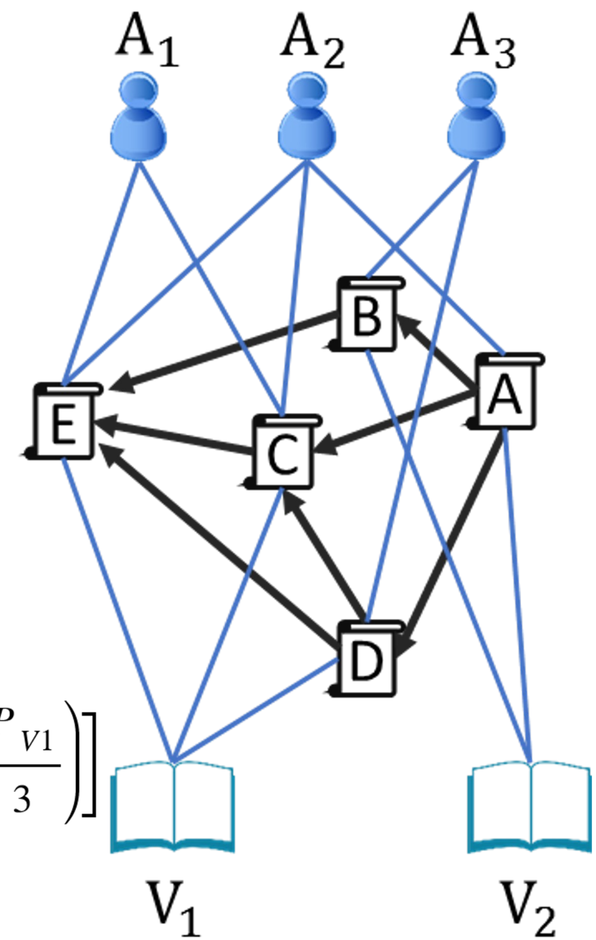
1. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.

Heterogeneous Networks

Applications

P-Rank[1]

- **Differentiate citations** based on citing papers, journals, authors
- Defines inter- and intra-graph **walks on heterogeneous network**
- **Author scores** based on their papers
- **Venue scores** based on their papers
- **"Random" Jump Vector** based on above, **run PageRank iteration**
- **Repeat until convergence**



$$P(C) = a\left[\frac{P(A)}{3} + \frac{P(D)}{2} + \frac{P(E)}{5}\right] + (1-a)\left[b\left(\frac{P_{A1}}{2} + \frac{P_{A2}}{3}\right) + c\left(\frac{P_{V1}}{3}\right)\right]$$

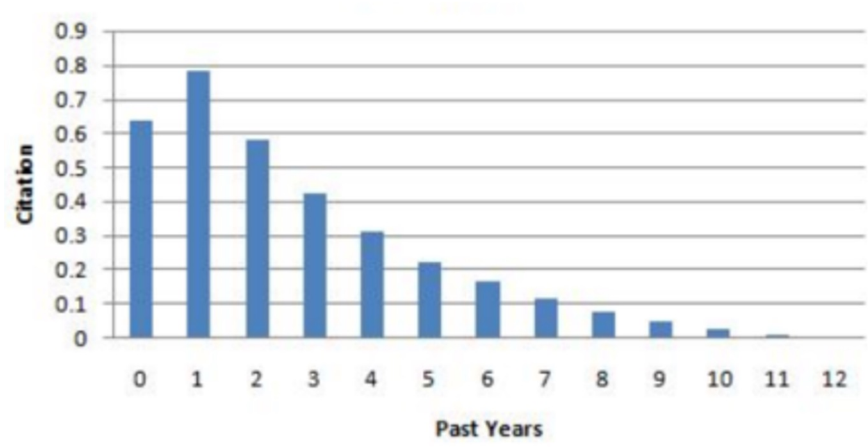
1. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.

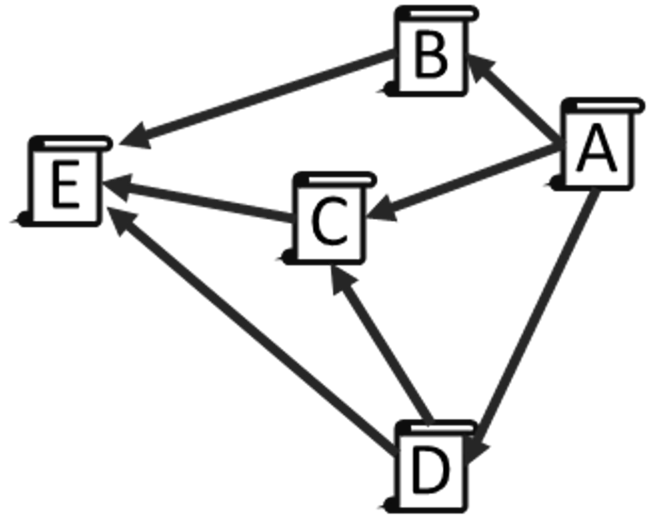
Heterogeneous Networks

Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- Most citations **made to papers published 1-2 years prior**
  - Hence, recently published papers are more important
  - Use **exponential weight** for paper age





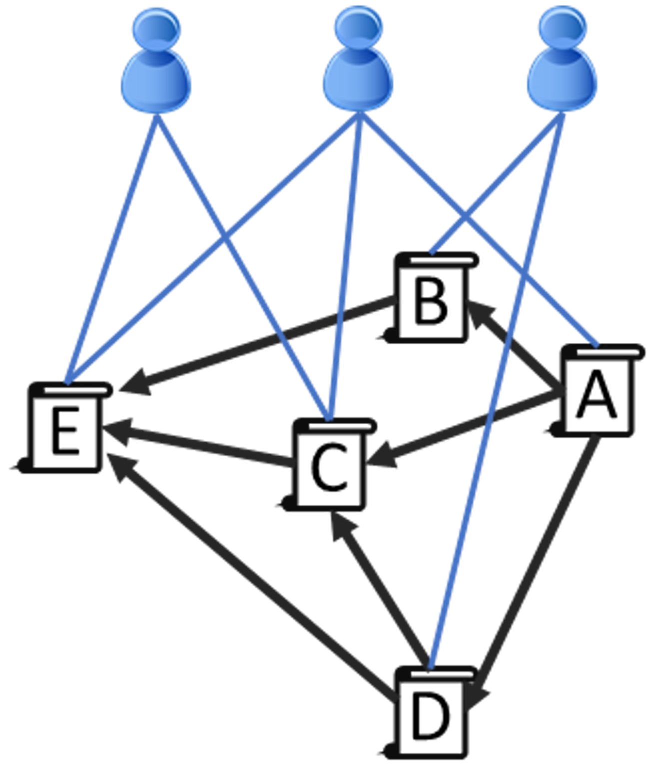
1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Network of papers and authors - mutual reinforcement between them
- M: authorship matrix, M[i,j]=1 iff paper j written by author i, else 0



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

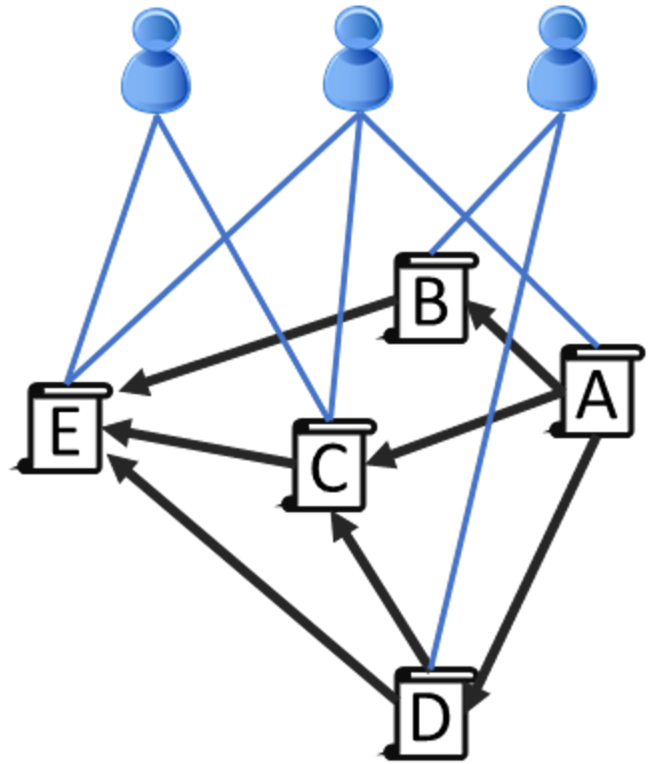
Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Repeat until convergence

$$FR(a_i) = \sum_j M[i,j] FR(p_j)$$

$$FR(p_i) = \alpha \sum_j S[i,j] FR(p_j) + \beta \sum_j M^T[i,j] FR(a_j) + \gamma e^{-\rho(T_C - T_i)}$$



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

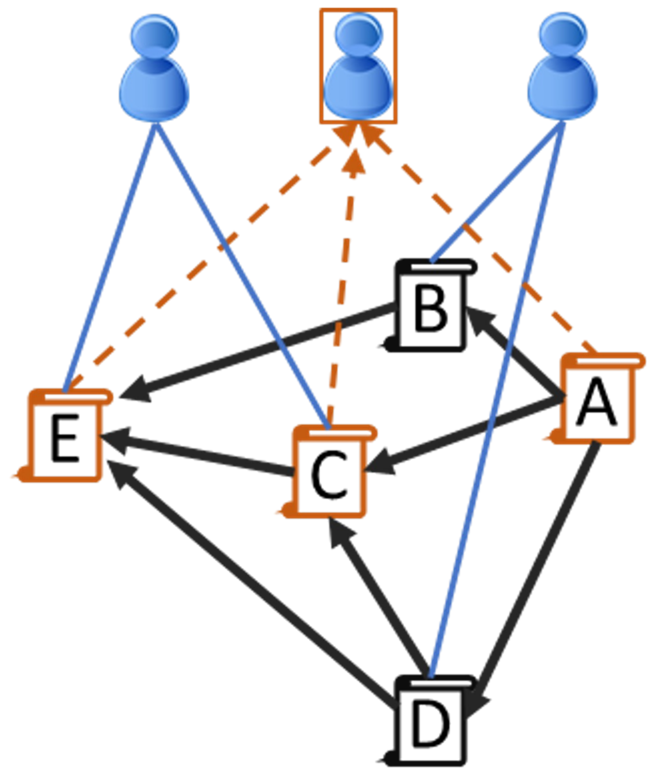
Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Repeat until convergence

$$\boxed{FR(a_i) = \sum_j M[i,j]FR(p_j)}$$

$$FR(p_i) = \alpha \sum_j S[i,j]FR(p_j) + \beta \sum_j M^T[i,j]FR(a_j) + \gamma e^{-\rho(T_C - T_i)}$$



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

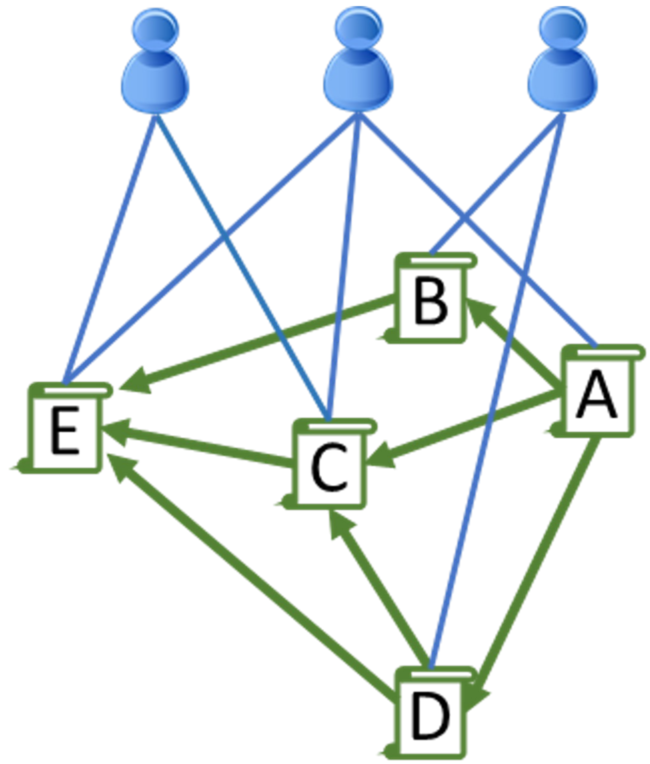
Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Repeat until convergence

$$FR(a_i) = \sum_j M[i,j] FR(p_j)$$

$$FR(p_i) = \boxed{\alpha \sum_j S[i,j] FR(p_j)} + \beta \sum_j M^T[i,j] FR(a_j) + \gamma e^{-\rho(T_C - T_i)}$$



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

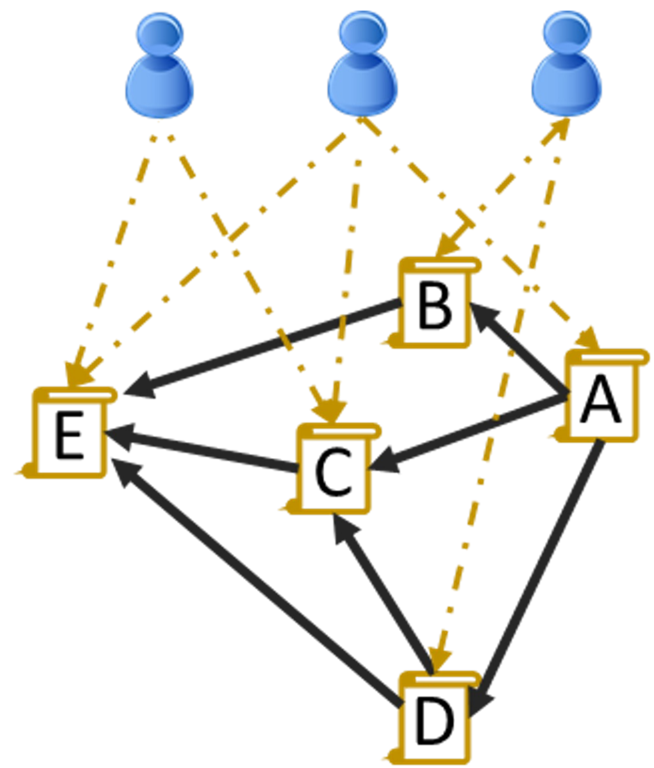
Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Repeat until convergence

$$FR(a_i) = \sum_j M[i,j] FR(p_j)$$

$$FR(p_i) = \alpha \sum_j S[i,j] FR(p_j) + \beta \sum_j M^T[i,j] FR(a_j) + \gamma e^{-\rho(T_C - T_i)}$$



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Heterogeneous Networks

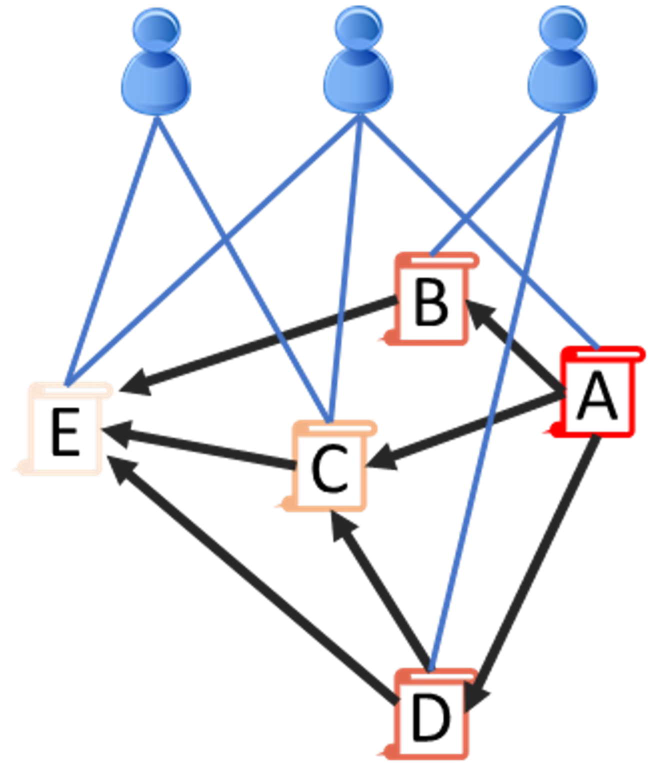
Applications

FutureRank[1]

- Goal: **predict PR scores in future graph**
- **"Good research is done by good researchers"**
- Repeat until convergence

$$FR(a_i) = \sum_j M[i,j] FR(p_j)$$

$$FR(p_i) = \alpha \sum_j S[i,j] FR(p_j) + \beta \sum_j M^T[i,j] FR(a_j) \boxed{+ \gamma e^{-\rho(T_C - T_i)}}$$



1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.

Classification II

Computational Model

Citation Count

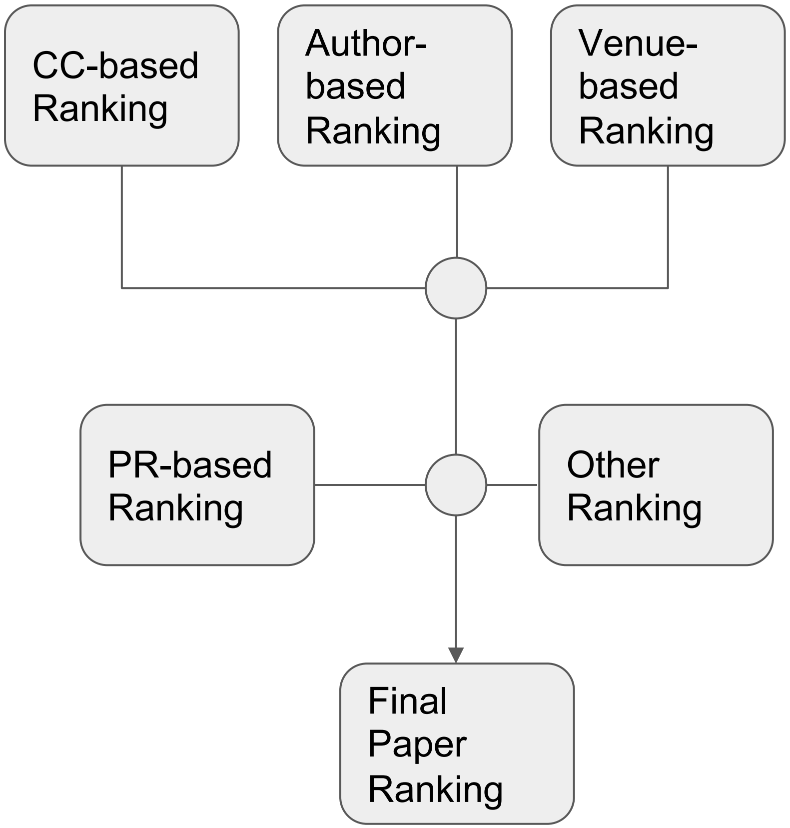
PageRank

Heterogeneous Networks

Ensemble Methods

- Calculate any number of different scores based on the above
- Combine them through some operator
- Most methods in KDD' cup 2016

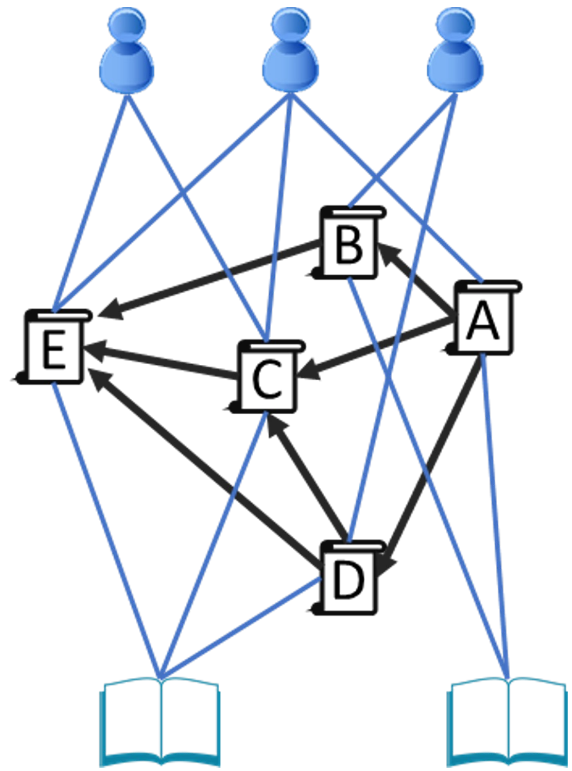
Other Approaches



Ensemble Methods
Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and "dominant group" avg
- Repeat ~ 5 times



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup. 2016.

Ensemble Methods

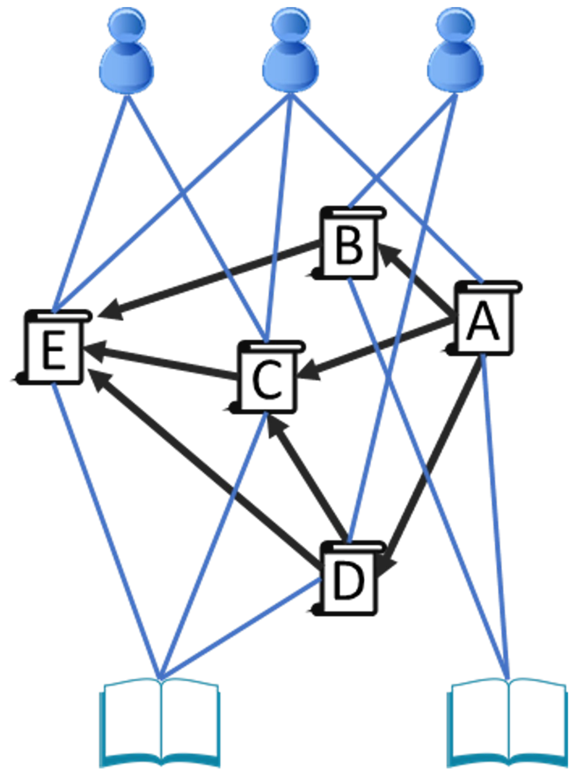
Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and "dominant group" avg
- Repeat ~ 5 times

$$P_i = \frac{\alpha(inDeg/maxInDeg) + \beta(outDeg/MaxOutDeg)}{1+\alpha}$$

$$P_C = \frac{\alpha(2/3) + \beta(1/3)}{1+\alpha}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup. 2016.

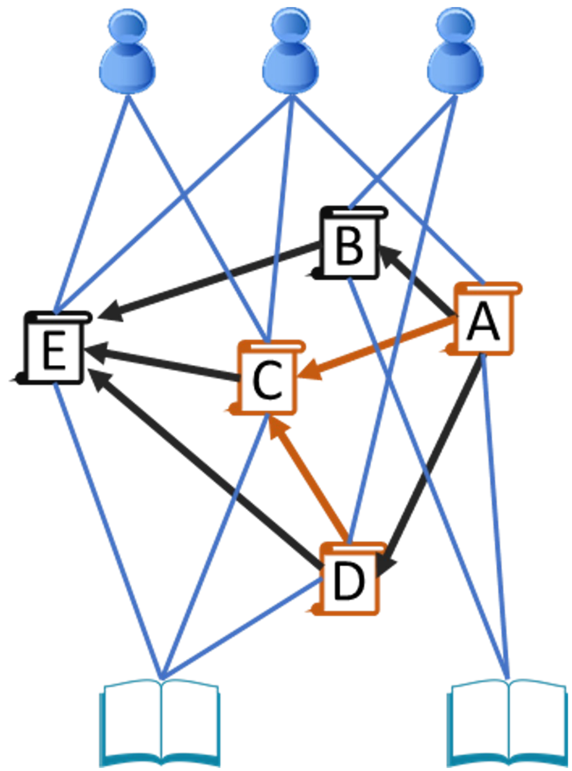
Ensemble Methods

Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and "dominant group" avg
- Repeat ~ 5 times

$$P(C) = \frac{P_A + P_D}{2}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup. 2016.

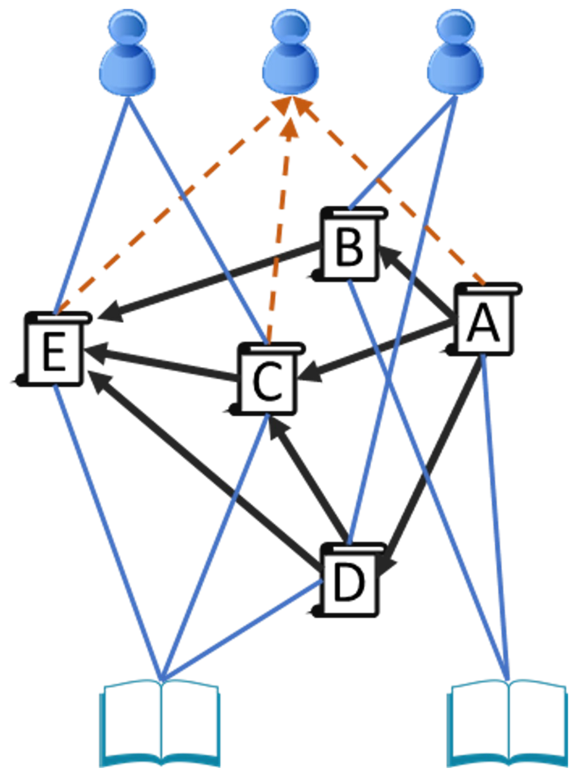
Ensemble Methods

Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and “dominant group” avg
- Repeat ~ 5 times

$$A_2 = \frac{P_C + P_E + P_A}{3}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner’s solution for wsdm cup 2016. WSDM Cup. 2016.

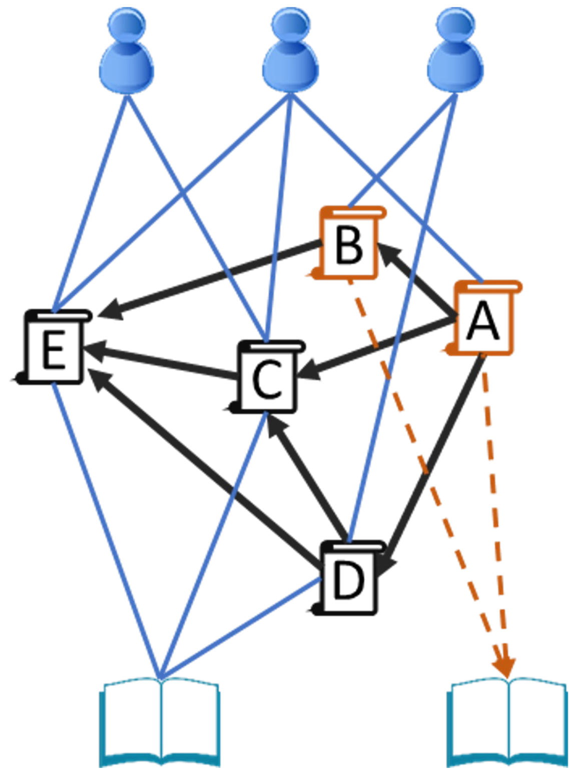
Ensemble Methods

Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and "dominant group" avg
- Repeat ~ 5 times

$$V_2 = \frac{P_A + P_B}{2}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup. 2016.

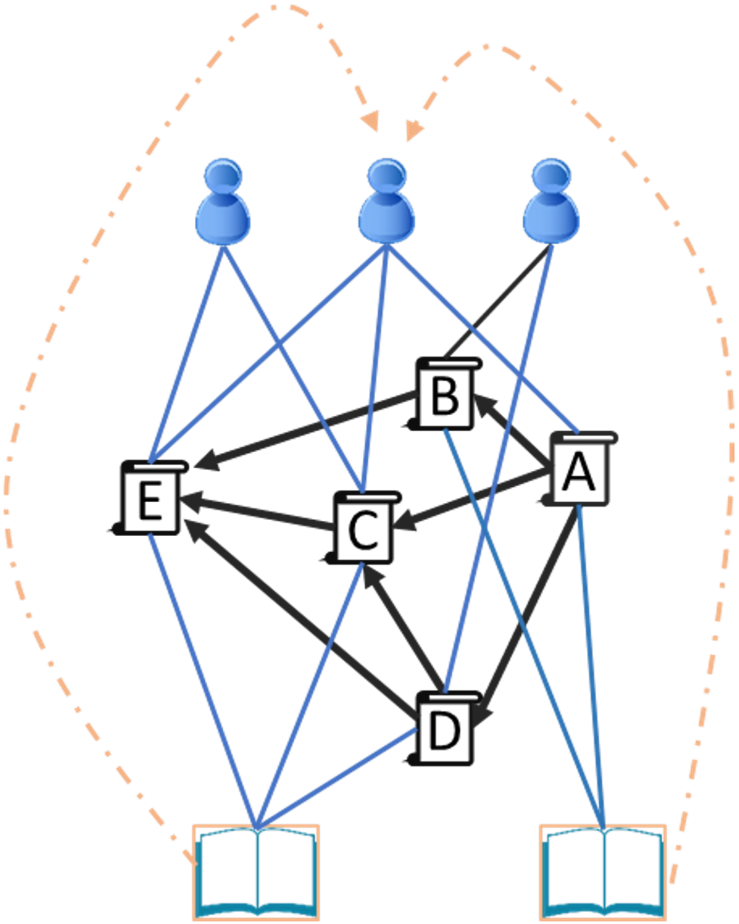
Ensemble Methods

Applications

WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and “dominant group” avg
- Repeat ~ 5 times

$$A'_2 = \frac{A_2 + (V_1 + V_2)/2}{2}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner’s solution for wsdm cup 2016. WSDM Cup. 2016.

Ensemble Methods

Applications

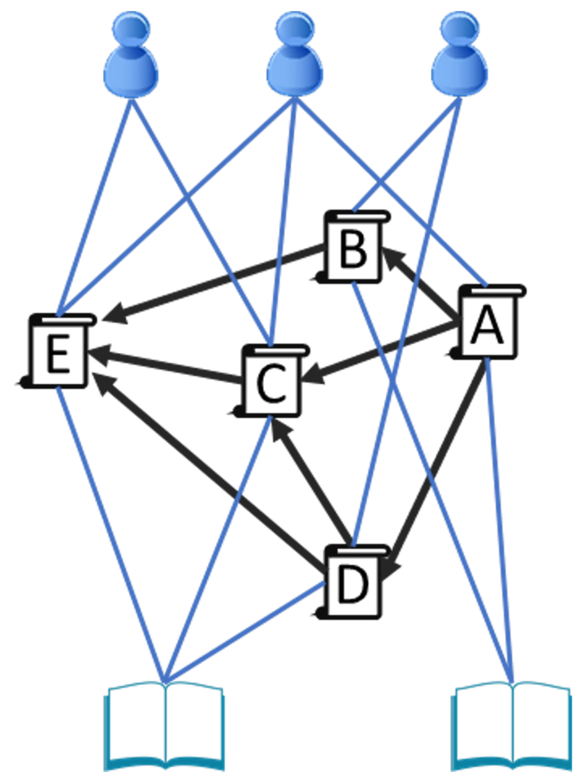
WSDM cup 2016 winner[1]

- Multiple **bipartite graphs**
- **Initialize**: linear combination of **citations and references**
- **Propagate** paper scores
  - Papers <= avg score of citing papers
  - Authors <= avg score of their papers
  - Venues <= avg score of their papers
- **Refine author** scores
  - Avg of previous step score based on the venues they publish in
- Apply **voting strategy**
  - Avg of initial score and "dominant group" avg
- Repeat ~ 5 times

$$P_V(C) = V_1\ /1$$
$$P_A(C) = (A_1 A_2)/2$$

$$P(C) > P_C$$
$$P_A(C) > P_C$$
$$P_V(C) < P_C$$

$$P'_C = \frac{\frac{P(C) + P_A(C)}{2} + P_C}{2}$$



1. Feng MH, Chan K, Chen HY, Tsai MF, Yeh MY, Lin SD. An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup. 2016.

Classification Axis II: underlying computational model

Citation Count

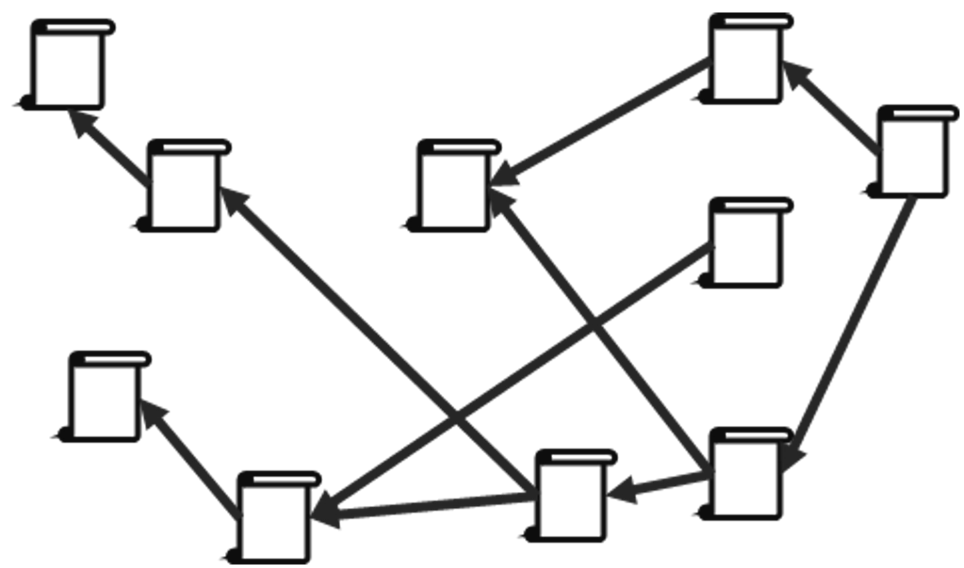
PageRank

Heterogeneous Networks

Ensemble Methods

Other Approaches

- Approaches not fitting the above
  - E.g., rescaling PageRank scores
  - using lengths of shortest citation paths
  - others



Other Approaches

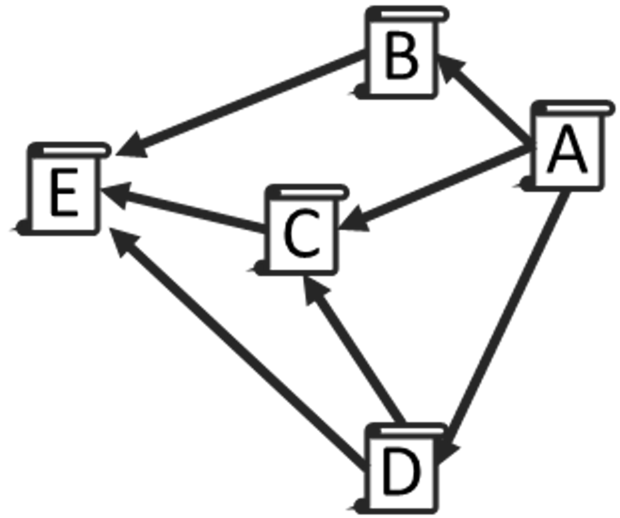
Example methods

Age-Rescaled PageRank[1]

- Goal: **debias age distribution** of highly ranked papers
- **Recalculate PageRank scores** based on other recently published papers

$$R(p_i) = \frac{PR(p_i) - \mu_i}{\sigma_i}$$

- Use papers j ∈ [i − Δp/2, i + Δp/2] to calculate avg and std dev
  - R(pi) < 0, **underperforms**
  - R(pi) < 0, **overperforms**
- Extension: field- & age-rescaled[2]



1. Mariani MS, Medo M, Zhang YC. Identification of milestone papers through time-balanced network centrality. Journal of Informetrics. 2016 Nov 1;10(4):1207-23.
2. Vaccario G, Medo M, Wider N, Mariani MS. Quantifying and suppressing ranking bias in a large citation network. Journal of informetrics. 2017 Aug 1;11(3):766-82.

Other Approaches

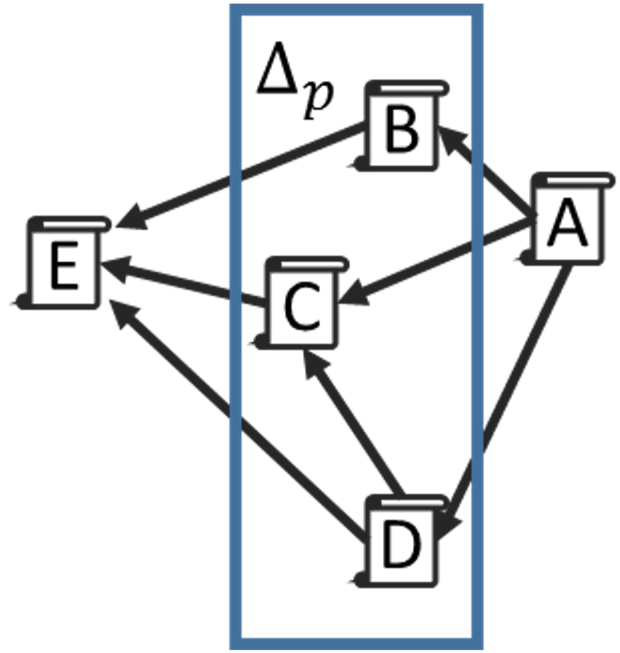
Example methods

Age-Rescaled PageRank[1]

- Goal: **debias age distribution** of highly ranked papers
- **Recalculate PageRank scores** based on other recently published papers

$$R(p_i) = \frac{PR(p_i) - \mu_i}{\sigma_i}$$

- Use papers j ∈ [i − Δp/2, i + Δp/2] to calculate avg and std dev
  - R(pi) < 0, **underperforms**
  - R(pi) > 0, **overperforms**
- Extension: field- & age-rescaled[2]



$$\mu_i = \frac{PR(C) + PR(B) + PR(D)}{3}$$

$$\sigma_i = \sqrt{\frac{(PR(B) - \mu_i)^2 + (PR(C) - \mu_i)^2 + (PR(D) - \mu_i)^2}{3}}$$

$$R(C) = \frac{PR(C) - \mu_i}{\sigma_i}$$

1. Mariani MS, Medo M, Zhang YC. Identification of milestone papers through time-balanced network centrality. Journal of Informetrics. 2016 Nov 1;10(4):1207-23.
2. Vaccario G, Medo M, Wider N, Mariani MS. Quantifying and suppressing ranking bias in a large citation network. Journal of informetrics. 2017 Aug 1;11(3):766-82.

Strengths and Weaknesses

General

Semantics

- PageRank-based models **translate** to researcher **behaviour**
  - Easier to understand
  - PageRank-based scores describe % of time spent on each paper or **probability of reaching a paper**
- Other methods lack these semantics
  - Some methods tuned based on some ground truth w/o providing any explainable semantics

Strengths and Weaknesses

General

Semantics

- PageRank-based models **translate** to researcher **behaviour**
  - Easier to understand
  - PageRank-based scores describe % of time spent on each paper or **probability of reaching a paper**
- Other methods lack these semantics
  - Some methods tuned based on some ground truth w/o providing any explainable semantics

Data usability

- Metadata-based approaches suffer from
  - Lesser availability
  - Data Cleaning issues

Strengths and Weaknesses

Popularity vs Influence

Time bias is inherent in Citation Count and PageRank

Some works place importance on **"predicting" rankings** based on **future citation counts or PageRank**

We examined effectiveness of different types of methods on this task

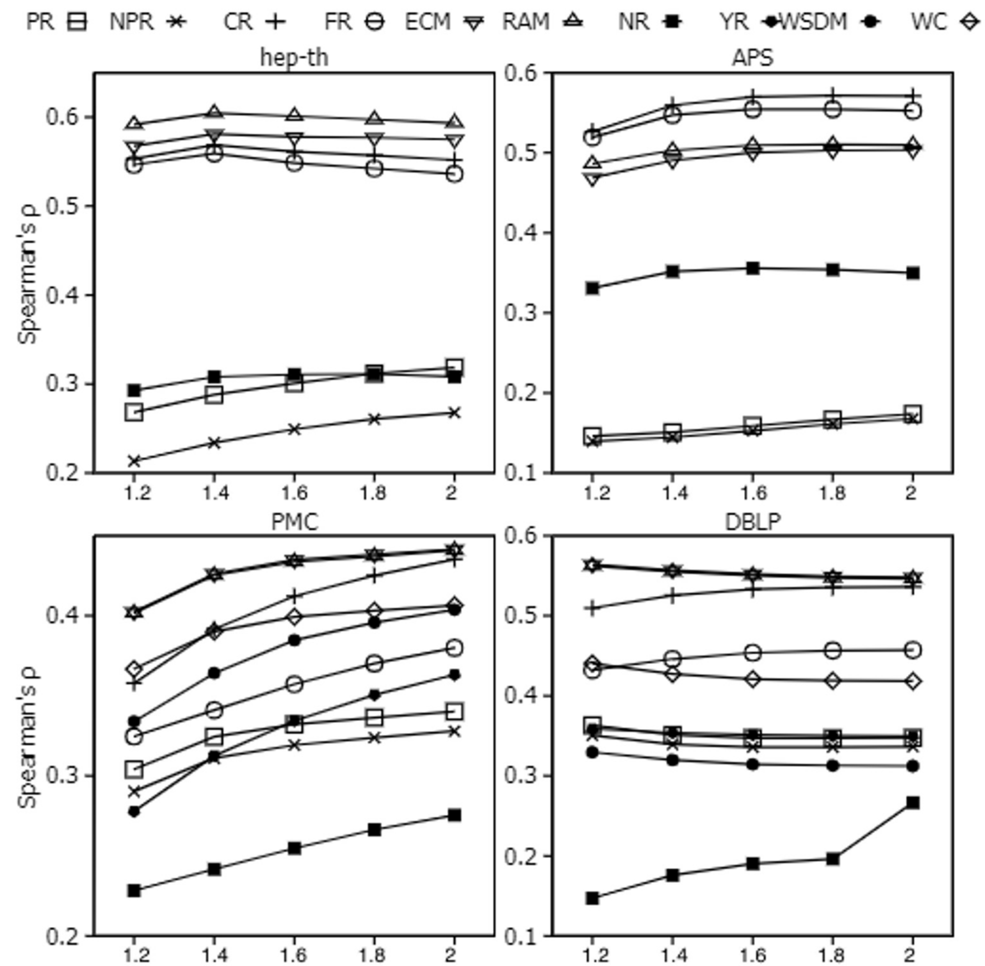
- Split dataset on time point t_s
- Rank papers based on examined method based on citation network up to t_s
- Compare ranking to
  - Future citation counts not counting old citations (Popularity)
  - Future citation counts considering all citations (Influence)

Strengths and Weaknesses

Popularity

Effectiveness on **Popularity**[1]

- Measure correlation of rankings to future citation counts (FCC)
- Time-aware methods perform best
  - **Citation age** most effective
  - Citation age **cannot capture cold start papers**
  - Paper age **cannot differentiate papers of same age**
  - Citation gap not as effective
- Metadata not **effective**



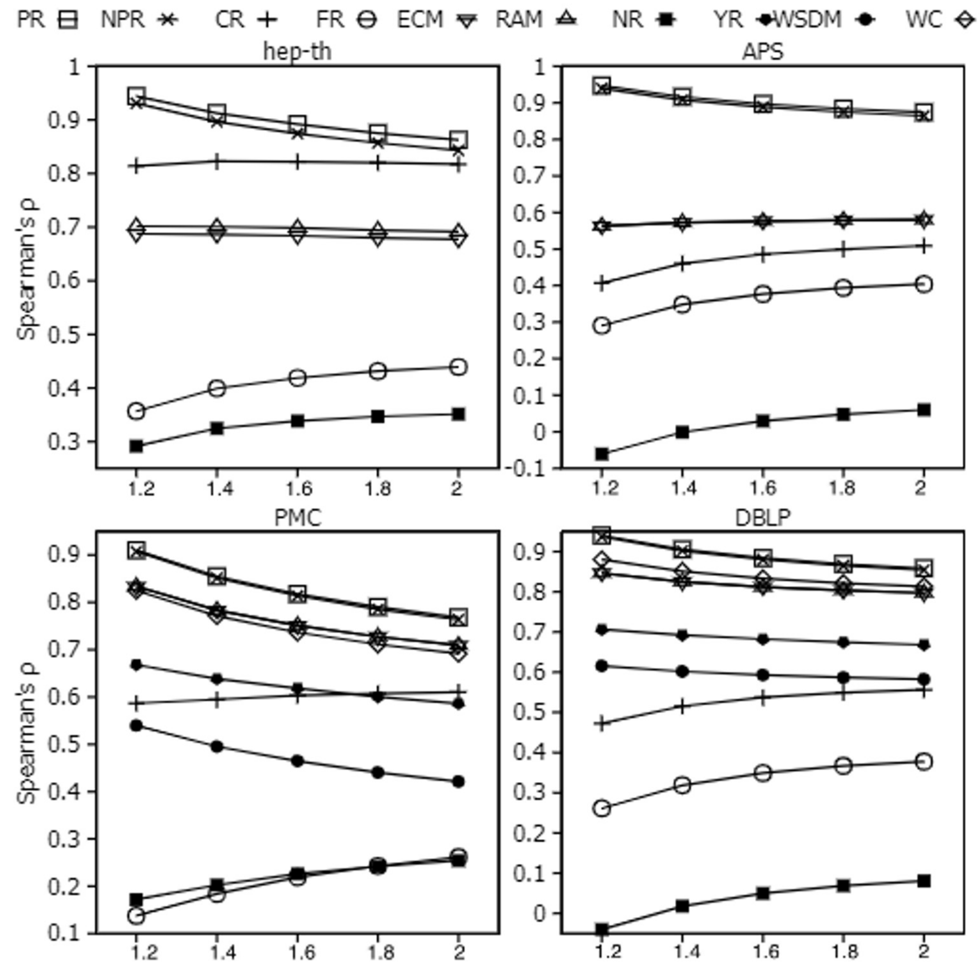
1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-based ranking of scientific publications: a survey and experimental evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13;33(4):1567-84.

Strengths and Weaknesses

Influence

Effectiveness on **Influence**[1]

- Measure correlation of rankings to overall PageRank - including future references (TPR)
- Traditional, time-independent methods are effective
- No particular benefit of ensemble / metadata-based methods



1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-based ranking of scientific publications: a survey and experimental evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13;33(4):1567-84.

Further Reading

1. Langville AN, Meyer CD. Google's PageRank and beyond. Princeton university press; 2011 Jul 1.
2. Chen P, Xie H, Maslov S, Redner S. Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics. 2007 Jan 1;1(1):8-15.
3. Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Information Processing & Management. 2008 Mar 1;44(2):800-10.
4. Hwang WS, Chae SM, Kim SW, Woo G. Yet another paper ranking algorithm advocating recent publications. InProceedings of the 19th international conference on World wide web 2010 Apr 26 (pp. 1117-1118).
5. Yao L, Wei T, Zeng A, Fan Y, Di Z. Ranking scientific publications: the effect of nonlinearity. Scientific reports. 2014 Oct 17;4(1):1-6.
6. Zhou J, Zeng A, Fan Y, Di Z. Ranking scientific publications with similarity-preferential mechanism. Scientometrics. 2016 Feb;106(2):805-16.
7. Krapivin M, Marchese M. Focused page rank in scientific papers ranking. InInternational Conference on Asian Digital Libraries 2008 Dec 2 (pp. 144-153). Springer, Berlin, Heidelberg.
8. Su C, Pan Y, Zhen Y, Ma Z, Yuan J, Guo H, Yu Z, Ma C, Wu Y. PrestigeRank: A new evaluation method for papers and journals. Journal of Informetrics. 2011 Jan 1;5(1):1-3.
9. Yan E, Ding Y. Weighted citation: An indicator of an article's prestige. Journal of the American Society for Information Science and Technology. 2010 Aug;61(8):1635-43.
10. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In2011 ieee 11th international conference on data mining workshops 2011 Dec 11 (pp. 373-380). IEEE.
11. Yu PS, Li X, Liu B. Adding the temporal dimension to search-a case study in publication search. InThe 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05) 2005 Sep 19 (pp. 543-549). IEEE.
12. Wade AD, Wang K, Sun Y, Gulli A. Wsdm cup 2016: Entity ranking challenge. InProceedings of the ninth ACM international conference on web search and data mining 2016 Feb 8 (pp. 593-594).
13. Walker D, Xie H, Yan KK, Maslov S. Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment. 2007 Jun 14;2007(06):P06010.

Further Reading

1. Sayyadi H, Getoor L. Futurerank: Ranking scientific articles by predicting their future pagerank. InProceedings of the 2009 SIAM International Conference on Data Mining 2009 Apr 30 (pp. 533-544). Society for Industrial and Applied Mathematics.
2. Zhang F, Wu S. Ranking scientific papers and venues in heterogeneous academic networks by mutual reinforcement. InProceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries 2018 May 23 (pp. 127-130).
3. Yan E, Ding Y, Sugimoto CR. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. Journal of the american society for information science and technology. 2011 Mar;62(3):467-77.
4. Wang Y, Tong Y, Zeng M. Ranking scientific articles by exploiting citations, authors, journals, and time information. InTwenty-seventh AAAI conference on artificial intelligence 2013 Jun 30.
5. Bai X, Xia F, Lee I, Zhang J, Ning Z. Identifying anomalous citations for objective evaluation of scholarly article impact. PloS one. 2016 Sep 8;11(9):e0162364.
6. Jiang X, Sun X, Zhuge H. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. InProceedings of the 21st ACM international conference on Information and knowledge management 2012 Oct 29 (pp. 714-723).
7. Liu Z, Huang H, Wei X, Mao X. Tri-rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce. In2014 IEEE 26th International Conference on Tools with Artificial Intelligence 2014 Nov 10 (pp. 493-500). IEEE.
8. Klosik DF, Bornholdt S. The citation wake of publications detects nobel laureates' papers. PloS one. 2014 Dec 1;9(12):e113184.
9. Mariani MS, Medo M, Zhang YC. Identification of milestone papers through time-balanced network centrality. Journal of Informetrics. 2016 Nov 1;10(4):1207-23.
10. Liao H, Mariani MS, Medo M, Zhang YC, Zhou MY. Ranking in evolving complex networks. Physics Reports. 2017 May 19;689:1-54.

Further Reading

Our relevant works

1. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Ranking papers by their short-term scientific impact. In2021 IEEE 37th International Conference on Data Engineering (ICDE) 2021 Apr 19 (pp. 1997-2002). IEEE.
2. Kanellos I, Vergoulis T, Sacharidis D. Ranking Papers by Expected Short-Term Impact. In Predicting the Dynamics of Research Impact 2021 (pp. 89-121). Springer, Cham.
3. Chatzopoulos S, Vergoulis T, Kanellos I, Dalamagas T, Tryfonopoulos C. Artsim: improved estimation of current impact for recent articles. In Adbis, tpdl and eda 2020 common workshops and doctoral consortium 2020 Aug 25 (pp. 323-334). Springer, Cham.
4. Chatzopoulos S, Vergoulis T, Kanellos I, Dalamagas T, Tryfonopoulos C. Further improvements on estimating the popularity of recently published papers. Quantitative Science Studies. 2021:1-36.
5. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-based ranking of scientific publications: a survey and experimental evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13;33(4):1567-84.
6. Vergoulis T, Chatzopoulos S, Kanellos I, Deligiannis P, Tryfonopoulos C, Dalamagas T. Bip! finder: Facilitating scientific literature search by exploiting impact-based ranking. InProceedings of the 28th ACM International Conference on Information and Knowledge Management 2019 Nov 3 (pp. 2937-2940).

Part C: Applications & Discussion

Thanasis Vergoulis (ATHENA RC, Greece)

Open SKGs are making it possible

- Open SKGs are catalysing research impact assessment applications.
- Ten years ago the coverage was extremely low.
- Important factor: the popularity of Open Science initiatives
  - Open citations
  - Open abstracts
  - Open SKGs

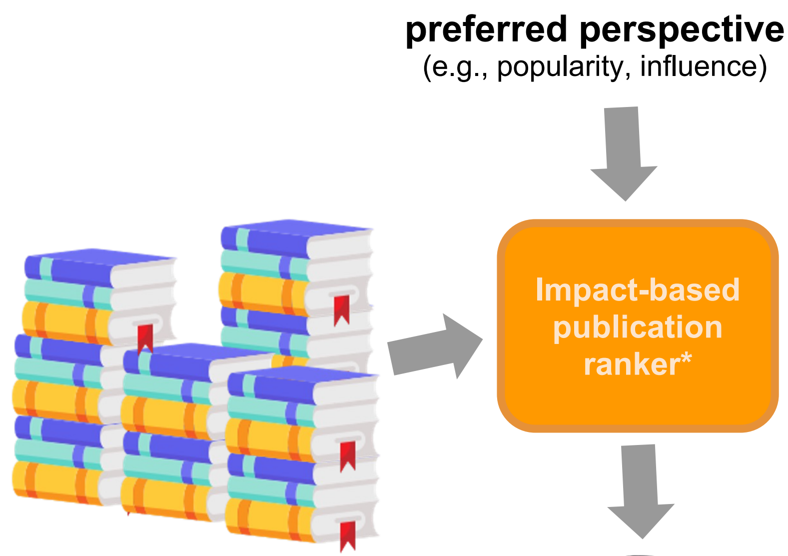
Real-world applications

- Impact measures for publications have various **real-world applications**.
    - Not restricted around publications assessment
- Example 1: **Literature reading prioritisation** (the traditional use)
    - Leverage impact measures to prioritise reading
    - Most common case: combine impact measures with keyword relevance scores
- Example 2: **Researcher assessment**
    - Evaluate the academic performance of a researcher according to the impact of their publications (and beyond!)
- Example 3: **Monitoring trends in research topics**
    - Take advantage of the cumulative impact of research topics to identify trends in their popularity

Literature reading prioritisation

The concept

- Delving into a field is tedious
- Extremely large number of published research works
- Existence of low-quality (even erroneous) works
- Different reads according to user / application (recall the experienced researcher Vs. student example)



* ranks papers according to the proper impact indicator (e.g., PageRank for influence, AttRank for popularity)

Literature reading prioritisation

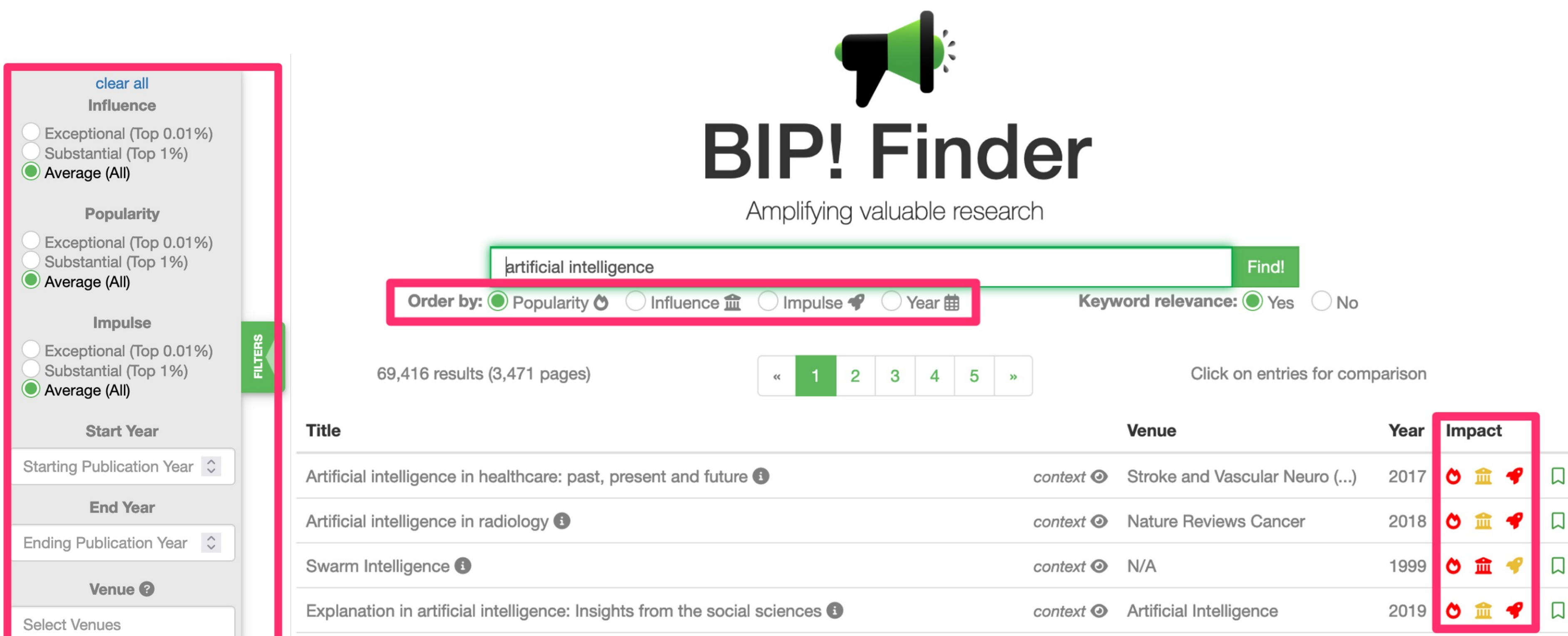
The prototype

https://bip.imsi.athenarc.gr/

Popularity = current attention
- RAM & AttRank

Influence = long-term importance
- CC & PageRank

Impulse = initial impact during “incubation phase”
- “incubation” CC (based on first 3y after publications)



Literature reading prioritisation

The resources (datasets, codes, APIs)

BIP! Vision: a set of services & resources to offer a ***multi-dimensional view of publications impact***

- BIP! DB: https://doi.org/10.5281/zenodo.4386934
- BIP! API: https://bip-api.imsi.athenarc.gr/documentation
- BIP! Ranker: https://github.com/athenarc/Bip-Ranker
- BIP4COVID19: https://doi.org/10.5281/zenodo.3723281

214,529 views | 31,381 downloads

Relevant publications:

T. Vergoulis, I. Kanellos, C. Atzori, A. Mannocci, S. Chatzopoulos, S. La Bruzzo, N. Manola, P. Manghi: **BIP! DB: A Dataset of Impact Measures for Scientific Publications.** WWW (Companion Volume) 2021: 456-460

T. Vergoulis, S. Chatzopoulos, I. Kanellos, P. Deligiannis, C. Tryfonopoulos, T. Dalamagas: **BIP! Finder: Facilitating scientific literature search by exploiting impact-based ranking.** CIKM 2019: 2937-2940 (demo)

THE WEB CONFERENCE ACM

Researcher assessment

The concept

Researcher assessment process → DECISION

- **career advancements** (e.g., recruitment, promotion)
- **prof. achievements** (e.g., prizes, funds)

Scrutinise CVs → extremely tedious

Use of **"evaluation shortcuts"**

- Number of papers
- IF of journals
- Citations / h-index

??

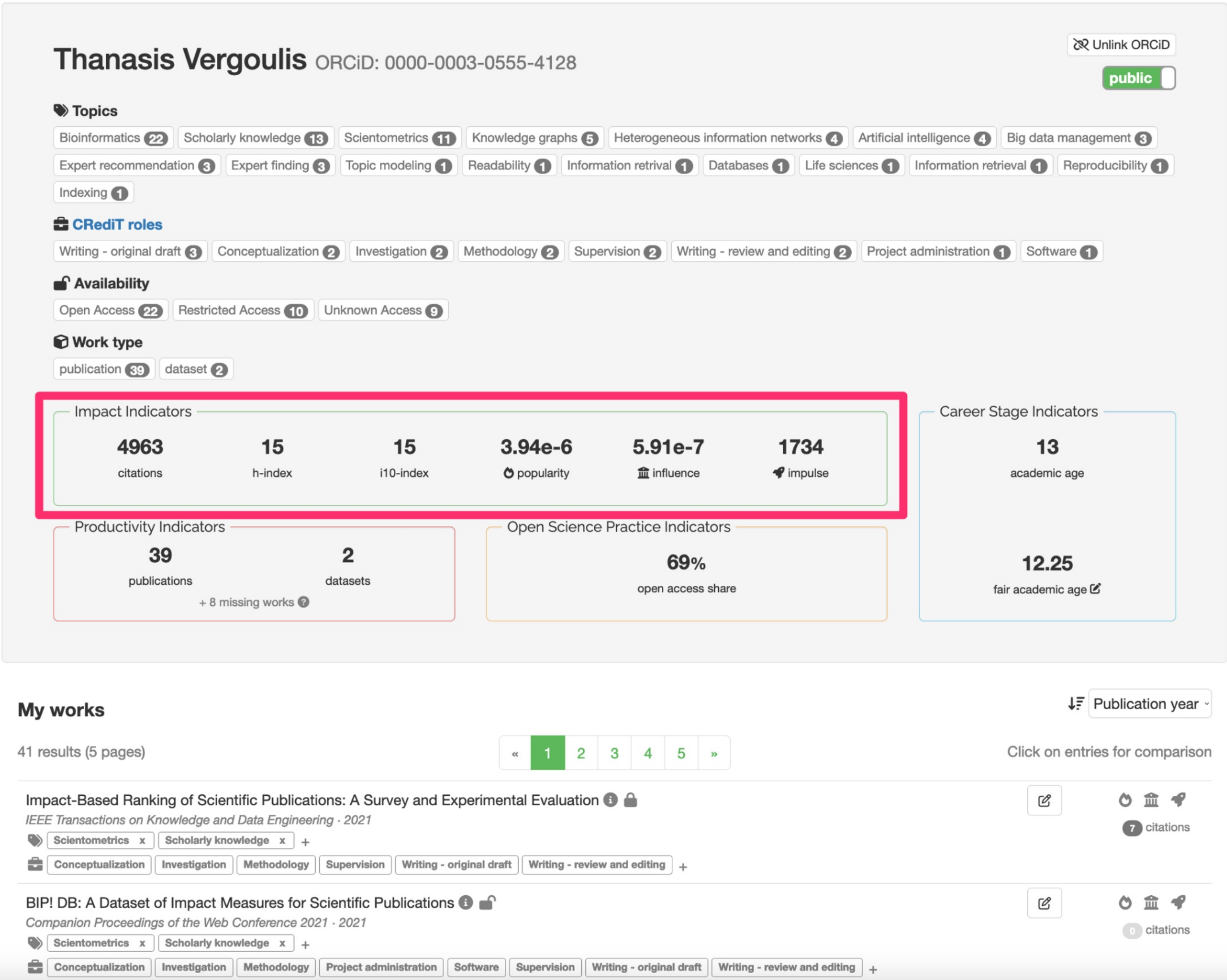
detailed impact indicators can help in this

many problems

- amplification of important problems (Matthew effect) / fostering bad practices
- **not capturing the full spectrum of aspects of research impact**
- oblivious of various types of research activities (datasets, software, OS practices)
- oblivious of different contribution roles
- misconceptions about the interpretation of the used indicators

Researcher assessment

Prototype



aggregated indicators capturing distinct impact aspects

provided details on the way of calculation, interpretations, misuses etc

additional indicators for other types of activity

adoption of contemporary RRA practices

option to manage their own profiles, adding CRediT roles for their contributions in the respective works

To appear in JCDL 2022

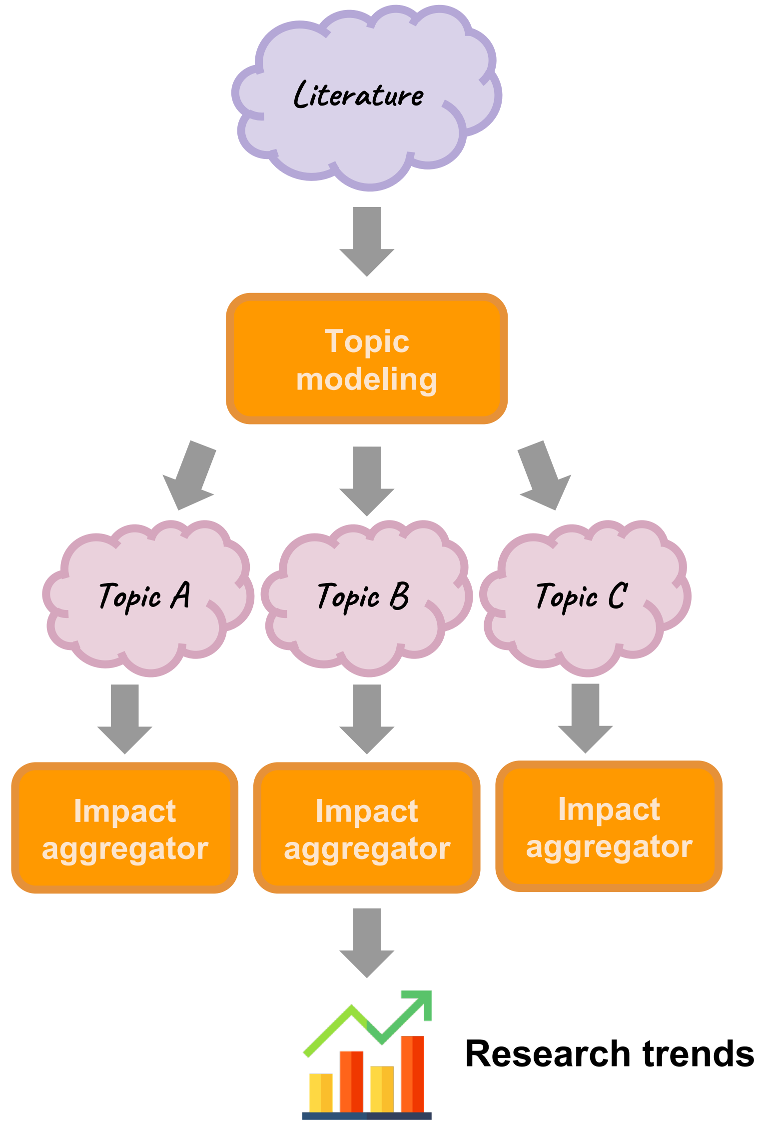
Monitoring trends in research topics

The concept

Officer in RFO (Research Funding Org.)

Research topics to fund???

Under-funded (new) topics??



Open challenges

- **Category A: Data quality & coverage in SKGs**
  - KG metadata that need to be enriched / cleaned
    - Examples: fields of study absent for most papers, author disambiguation
  - Full texts useful be available (at least in inv. Index form)
- **Category B: Improvements in indicators**
  - Multi-perspective field-weighted indicators
  - Impact propagation to other types of research output (e.g., datasets, software)
  - Incorporate citation semantics information
    - When using citations as proxies of impact, some citations may be irrelevant

Don’t forget to join us tomorrow:
https://sci-k.github.io

Thank you!

Ilias Kanellos - ilias.kanellos@athenarc.gr

Dimitris Sacharidis - dimitris.sacharidis@ulb.be - @dsachar

Thanasis Vergoulis - vergoulis@athenarc.gr - @vergoulis