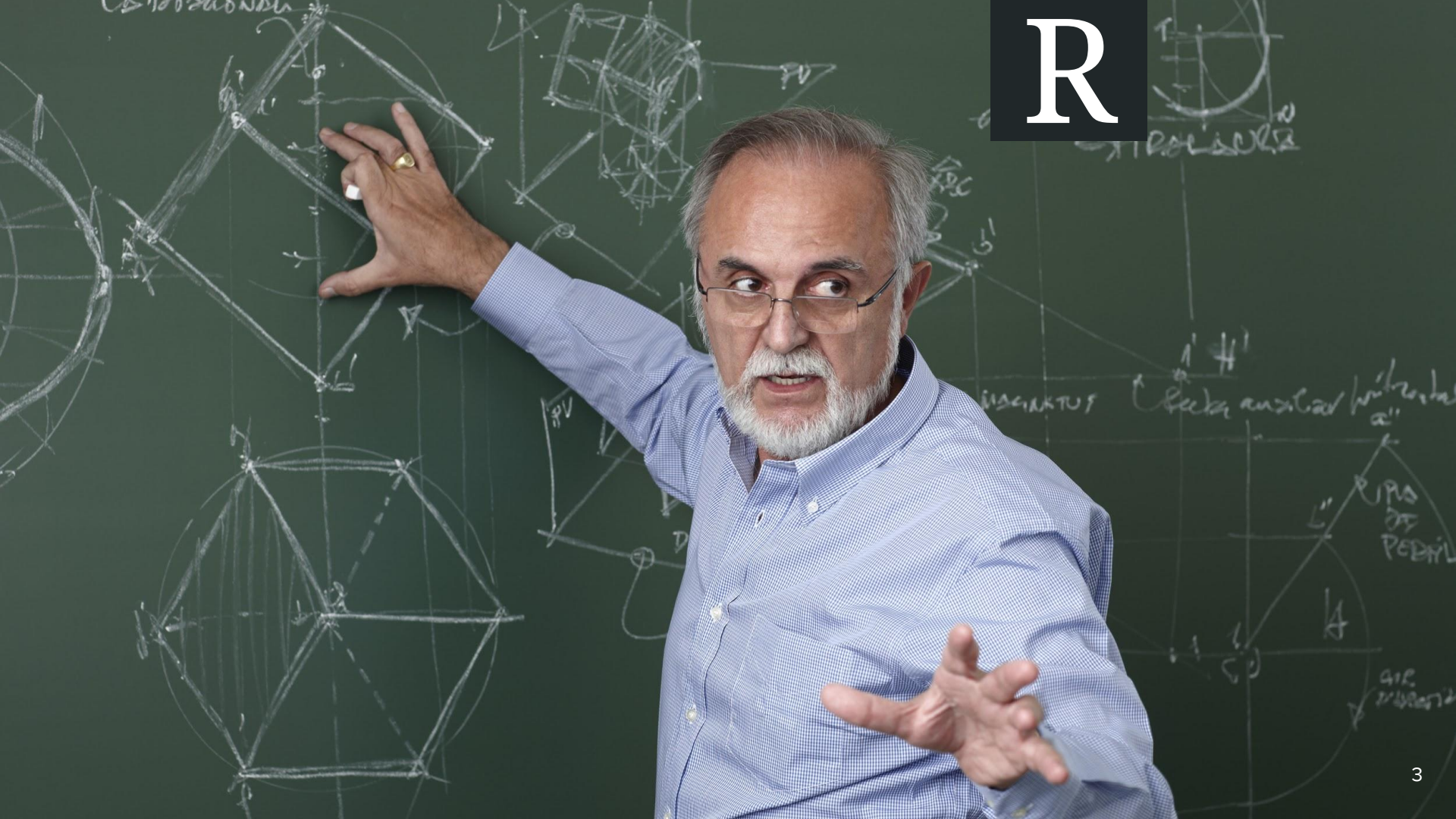
Reputation in Computer Science on a per Subarea Basis

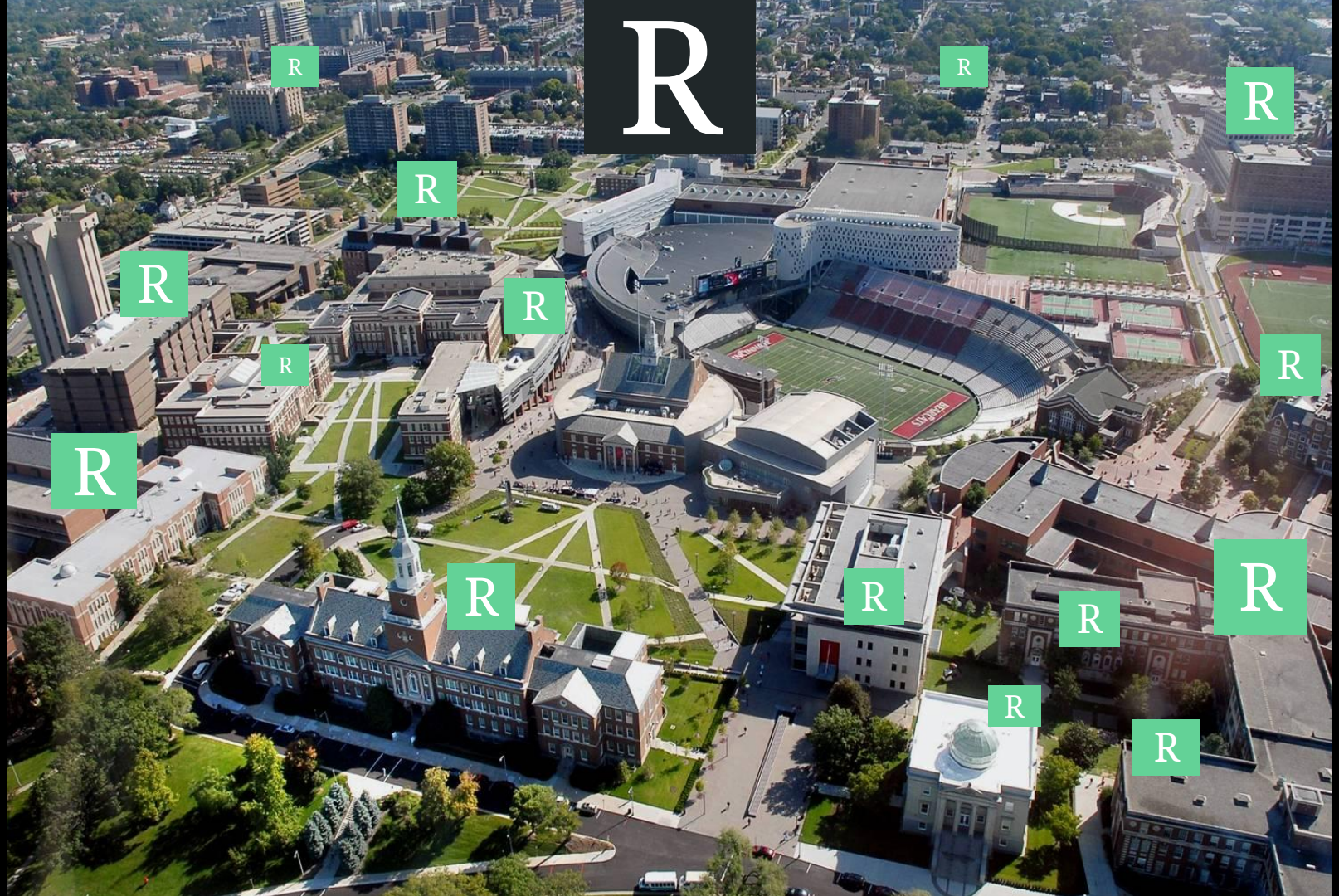
Alberto Hideki Ueda
Berthier Ribeiro-Neto and Nivio Ziviani

July 14, 2017

ueda@dcc.ufmg.br


R
2015无菌药品包装技术论坛
CNPPA
AGENDA
AGENDA


R


R
R
R
R
R
R
R
R
R
R
R
R
R
R
R
R
R

Motivation

Assessing the reputation of academic entities is an important task from different perspectives

37 Subareas in CS

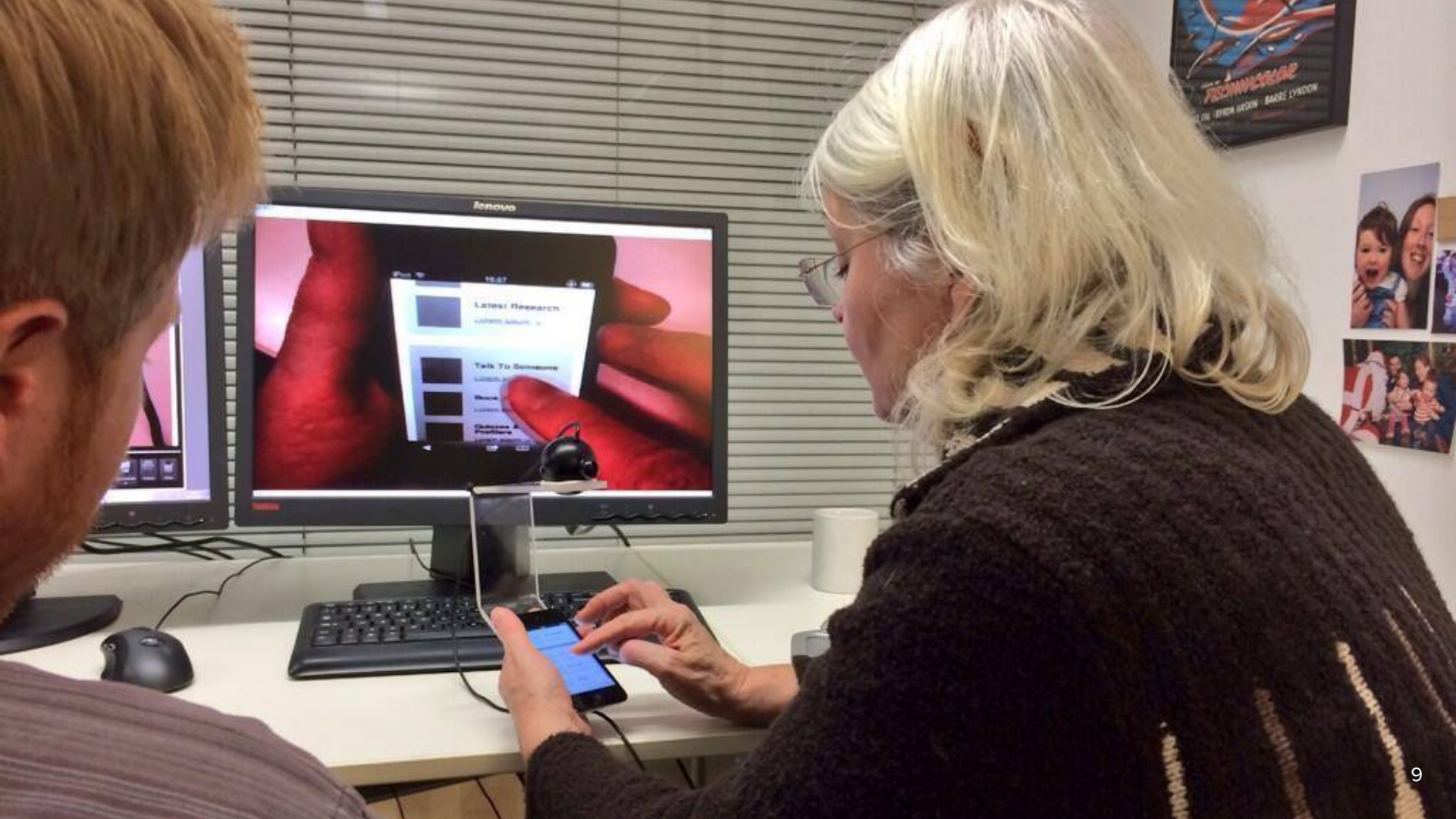
Microsoft Academic, 2017

CS Subareas	
Algorithm	Internet privacy
Artificial intelligence	Knowledge management
Bioinformatics	Machine learning
Cognitive science	Management science
Computational biology	Mathematical optimization
Computational science	Multimedia
Computer architecture	Natural language processing
Computer graphics	Operating system
Computer hardware	Operations research
Computer network	Parallel computing
Computer security	Pattern recognition
Computer vision	Programming language
Data mining	Real-time computing
Data science	Simulation
Database	Speech recognition
Distributed computing	Telecommunications
Embedded system	Theoretical computer science
Human-computer interaction	World Wide Web
Information retrieval	

Image Processing and Computer Vision
Average of **5 publications/year** in top venues

Operational Research and Optimization

Average of **5 publications/year** in top venues


lenovo
Latest Research
Talk To Someone

PROOF OF SPECTRAL THEOREM FOR SQUARE MATRICES

(from P. Lax "Linear Algebra")

Lemma 1: Let p & q be 2 polynomials with complex coefficients. Assume p & q have no common zero. Then, one can find 2 other polynomials a & b such that

$$ap + bq = 1.$$

Proof: p & q are given. Consider the set

$$\mathcal{P} = \{\alpha p + \beta q \mid \alpha, \beta \text{ polynomials}\}.$$

There is at least one non-zero polynomial in $\mathcal{P}$ of lowest degree. Call $d = \alpha p + \beta q$ such a polynomial.

. By the division algorithm, one can write $p = md + r$ where m and r are polynomials.

Since $r = p - md = p - m(\alpha p + \beta q)$, then $r \in \mathcal{P}$.

But the degree of r must be strictly less than the degree of d, which together with $r \in \mathcal{P}$ implies $r = 0$. Then, $p = md$, ie. d divides p. Similarly, d divides q.

. Now we show that d is a constant. If d is not of degree 0, then it has at least a root, which is also a root of p and q. Since we have assumed that p and q have no common roots, this is a contradiction.

. So $\alpha p + \beta q = d = \text{constant} \Rightarrow \frac{\alpha}{d} p + \frac{\beta}{d} q = 1$.

Lemma 2: Let p & q be as in lemma 1, and let A be a square matrix with complex entries. Denote by N_p, N_q and N_{pq} the nullspaces of $p(A)$, $q(A)$ and $p(A)q(A)$ respectively.

37 Subareas in CS

Microsoft Academic, 2017

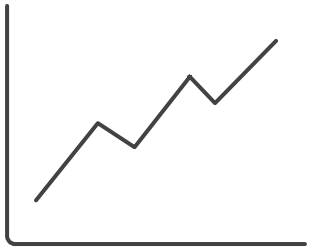
CS Subareas	
Algorithm	Internet privacy
Artificial intelligence	Knowledge management
Bioinformatics	Machine learning
Cognitive science	Management science
Computational biology	Mathematical optimization
Computational science	Multimedia
Computer architecture	Natural language processing
Computer graphics	Operating system
Computer hardware	Operations research
Computer network	Parallel computing
Computer security	Pattern recognition
Computer vision	Programming language
Data mining	Real-time computing
Data science	Simulation
Database	Speech recognition
Distributed computing	Telecommunications
Embedded system	Theoretical computer science
Human-computer interaction	World Wide Web
Information retrieval	


2015无菌药品包装技术论坛
CNPPA
AGENDA

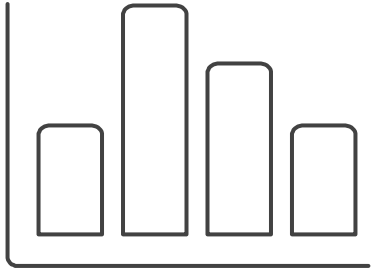

2015无菌药品包装技术论坛
CNPPA

Reputation in CS on a per Subarea Basis

How to **quantify** it?



How does it **vary** per CS subarea?



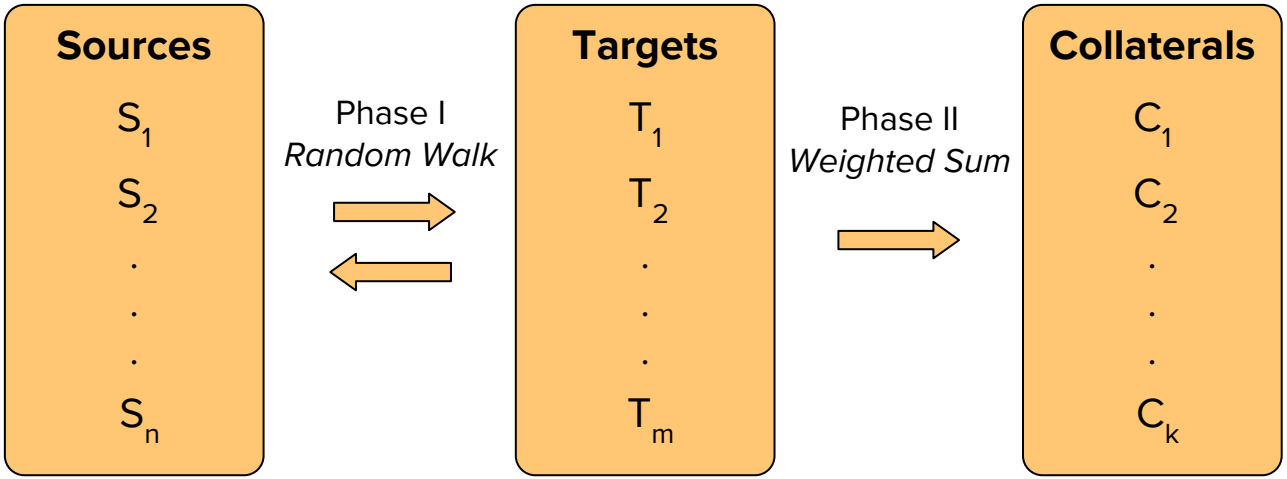
How does it **differ** in BR and US?

Our hypothesis

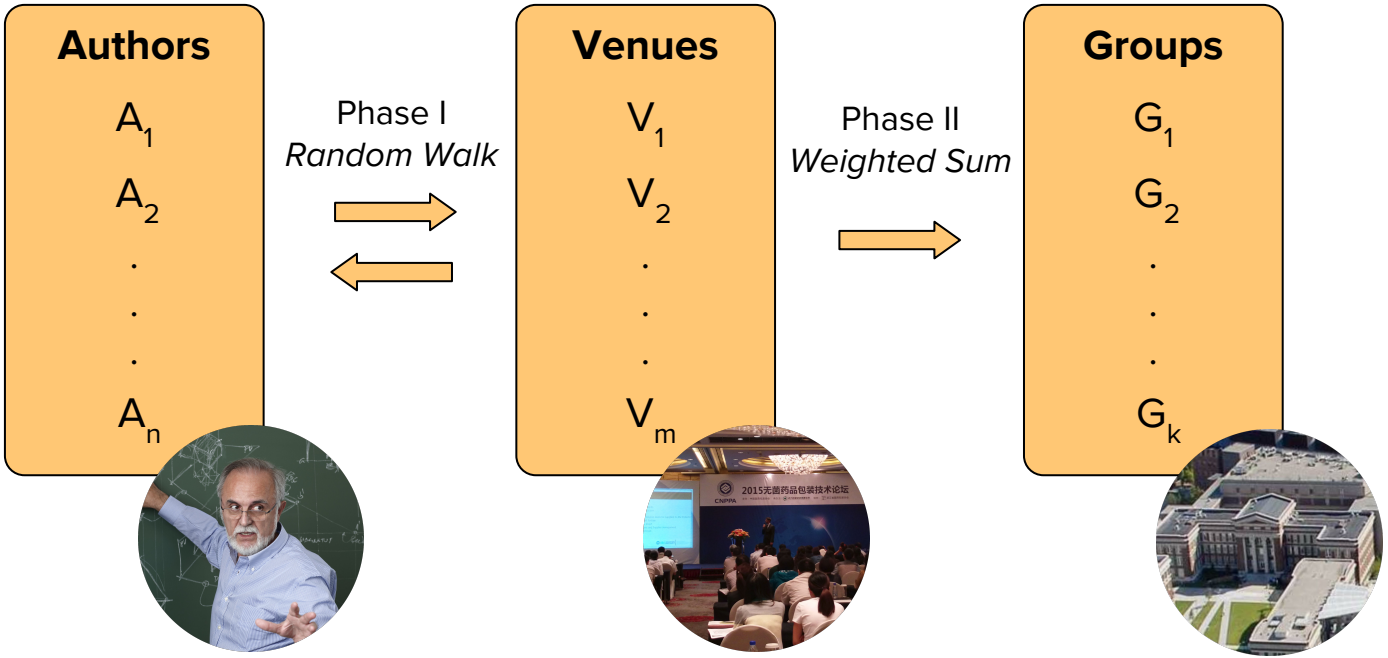
Using P-score on a per subarea basis we can improve rankings of academic entities.

Reputation Flows

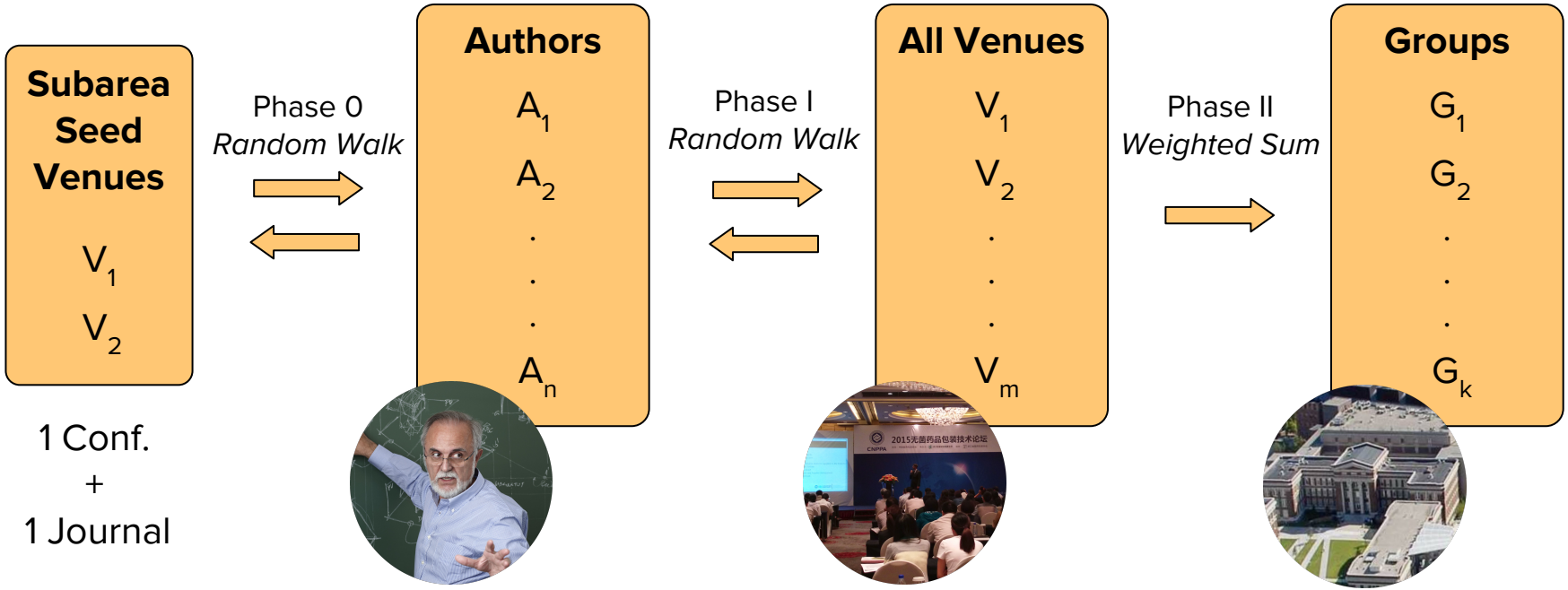
Reputation Flows



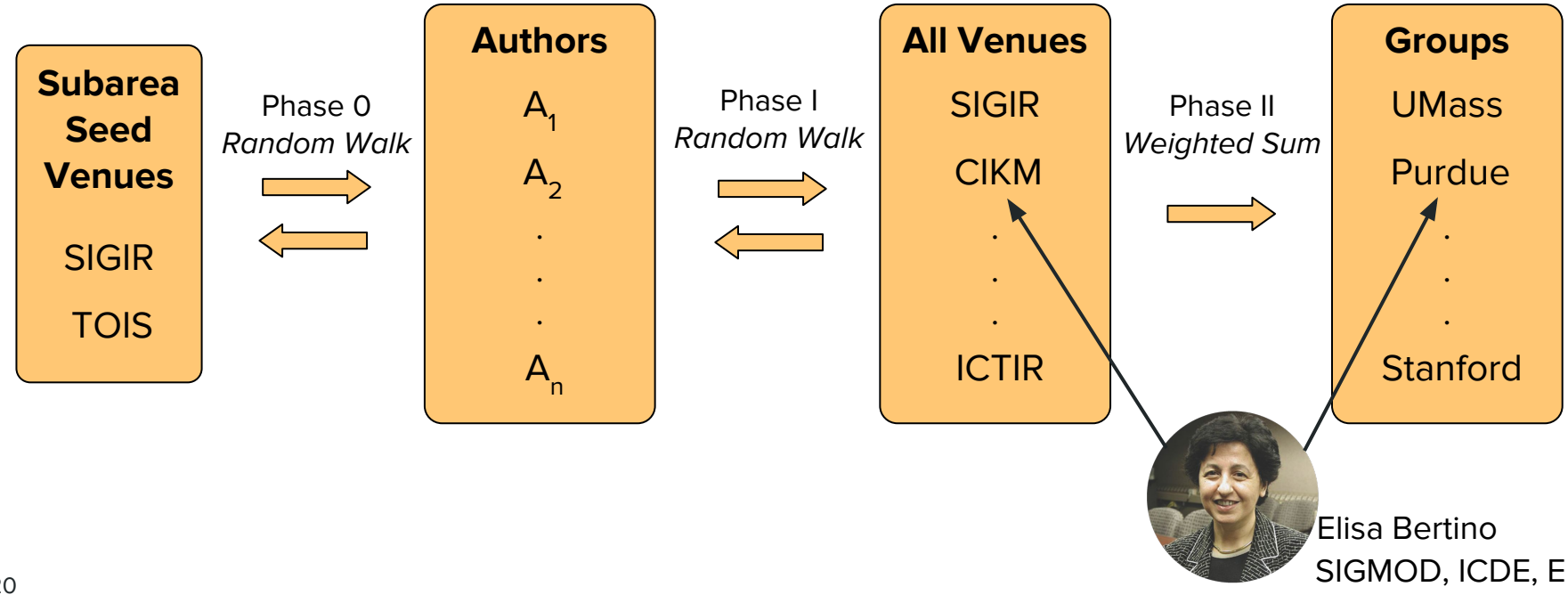
Reputation Flows



Standard P-score

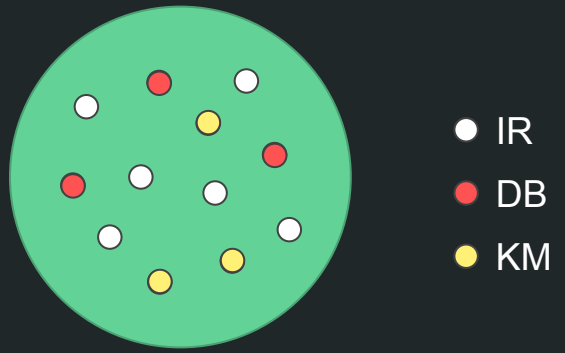


Standard P-score for IR



The Encroachment Problem

Publication venues possibly cover multiple subareas.



Normalized P-score for Venues

$$norm\text{-}P\text{-}score(v) = \frac{P\text{-}score(v)}{number\_of\_publications(v)}$$

$$V_s = \{v \in V \mid norm\text{-}P\text{-}score(v) > \alpha_s\}$$

v_1 v_2 ... v_n

$$\gamma(a,s) = \frac{1}{|\theta(a)|} \sum_{p \in \theta(a)} \begin{cases} \frac{1}{\log_2(y(0)-y(p)+2)} & \text{if } p \in V_s \\ 0 & \text{otherwise} \end{cases}$$

Factor γ of author a for subarea s

Set of publications of author a

Year of publication of paper p

Weighted P-score for Graduate Programs

$$P\text{-}score(g) = \sum_{a \in \phi(g)} \sum_{p \in \theta(a)} \frac{P\text{-}score(venue(p))}{number\_of\_\ authors(p)}$$

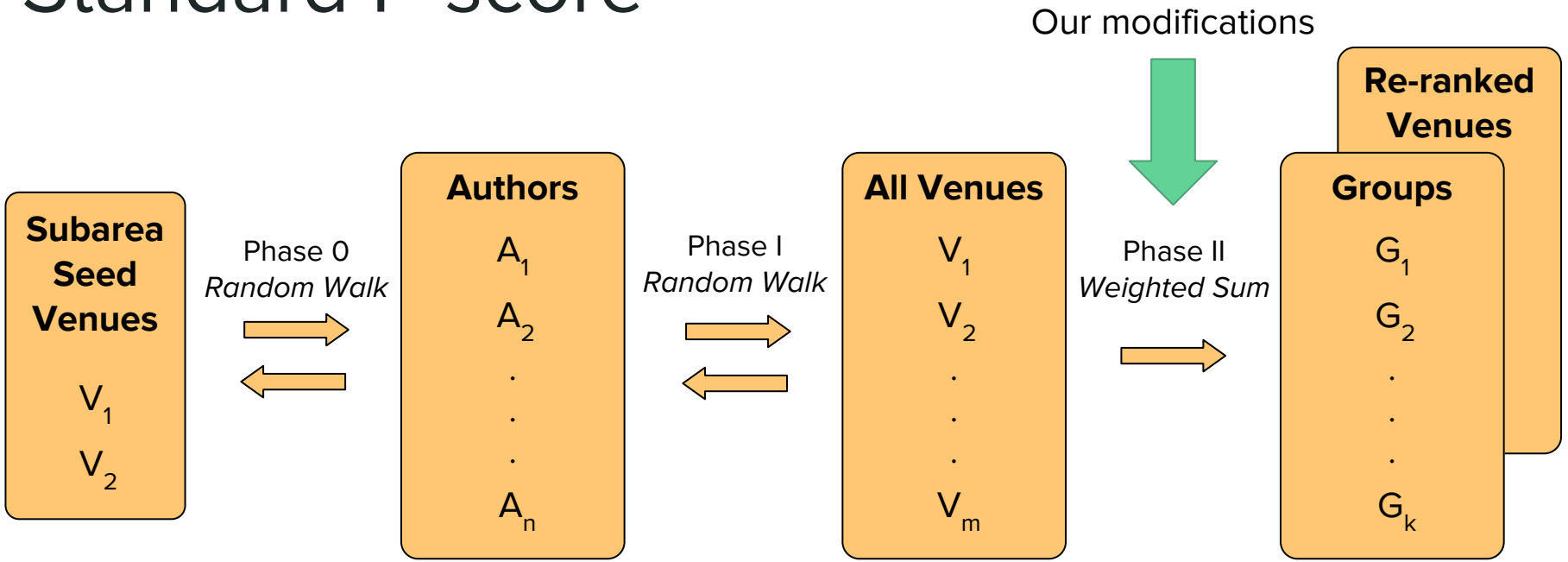
$$weighted\text{-}P\text{-}score(g, s) = \sum_{a \in \phi(g)} \gamma(a, s) \times \sum_{p \in \theta(a)} \frac{P\text{-}score(venue(p))}{number\_of\_authors(p)}$$

Set of researchers of group g

Factor γ of author a for subarea s

Set of publications of author a

Standard P-score



Experimental Setup

Dataset

Seed Venues

Venues Ground-truth

Dataset

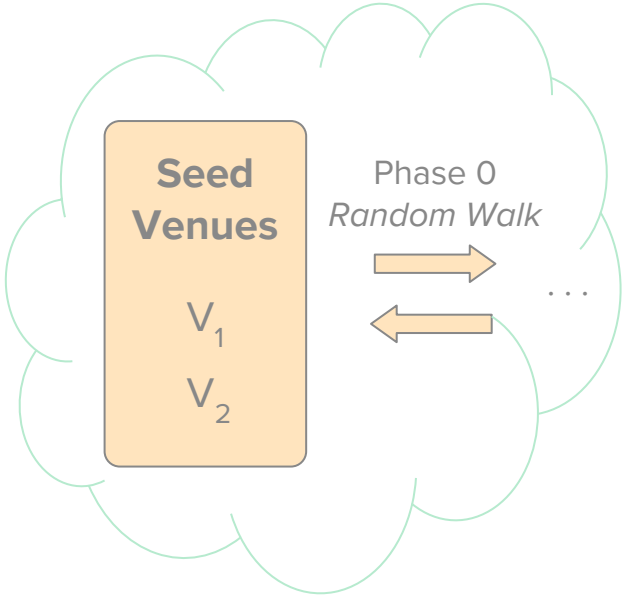
dblp
computer science bibliography

CS Rankings (beta)

Emery Berger
UMass

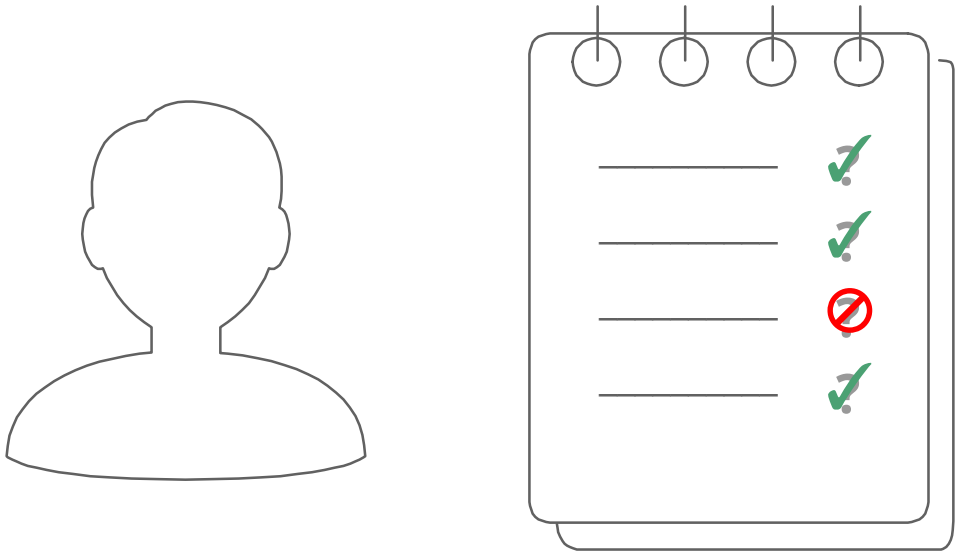
Attribute	Value
Number of Papers	2,931,849
Number of Authors	1,595,771
Number of Venues	5,765
Number of US Depts	126
Avg. number of professors per US Dept	42.4
Number of BR Depts	25
Avg. number of professors per BR Dept	47.8

Seed Venues

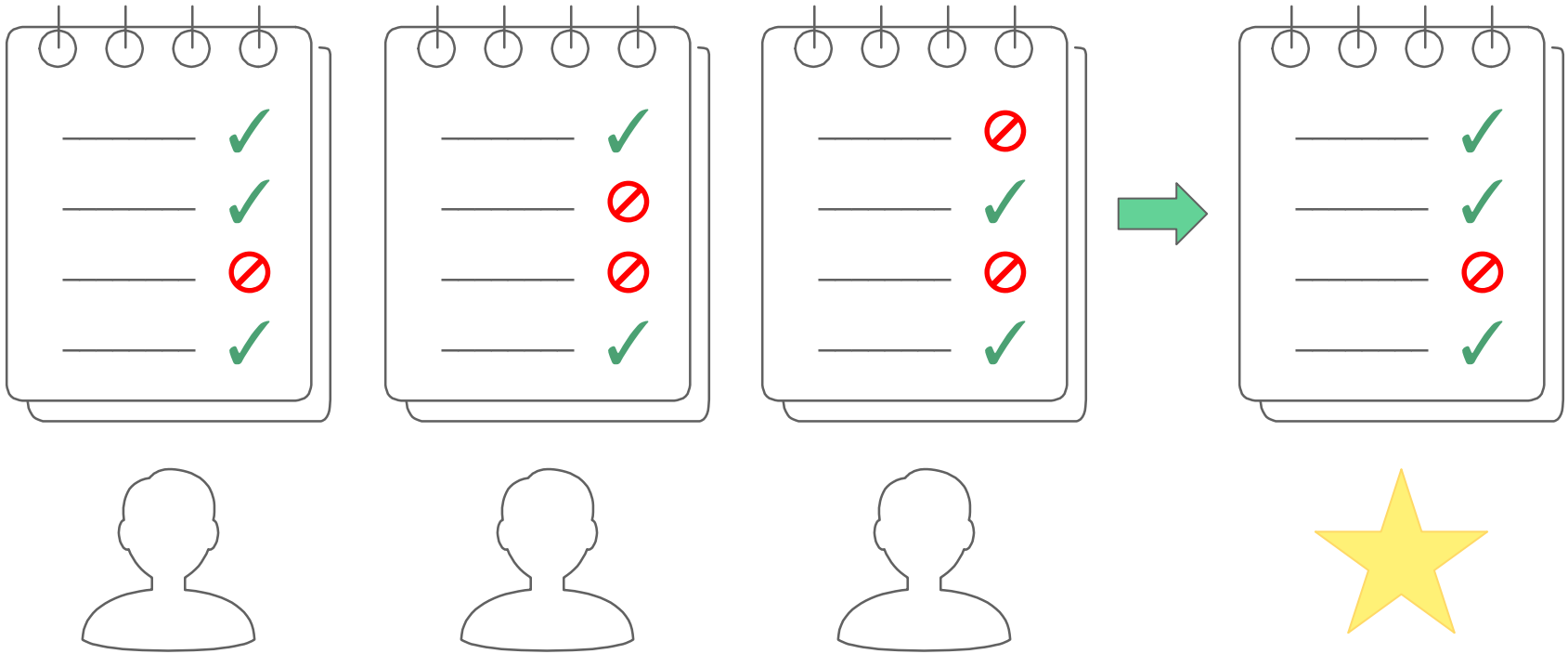


Subarea	Abbreviation	Seed venues
Algorithm	Alg	SODA, Algorithmica
Artificial intelligence	AI	IJCAI, AI
Bioinformatics	Bio	BIBM, Bioinformatics
Computer graphics	CG	SIGGRAPH, TCVG
Computer network	CN	INFOCOM, TON
Computer security	CS	CCS, TISSEC
Computer vision	CV	CVPR, IJCV
Data mining	DM	KDD, SIGKDD
Database	DB	SIGMOD, TODS
Distributed computing	DC	ICDCS, TPDS
Human-computer interaction	HCI	CHI, TOCHI
Information Retrieval	IR	SIGIR, TOIS
Machine learning	ML	ICML, JMLR
Natural language processing	NLP	EMNLP, COLING
Operating system	OS	SOSP, SIGOPS
Parallel computing	PC	IPPS, TPDS
Programming language	PL	PLDI, TOPLAS
Speech Recognition	SR	INTERSPEECH, TCOM
Theoretical computer science	TCS	STOC, SIAMCOMP
World Wide Web	WWW	WWW, WS

Venues Ground-Truth

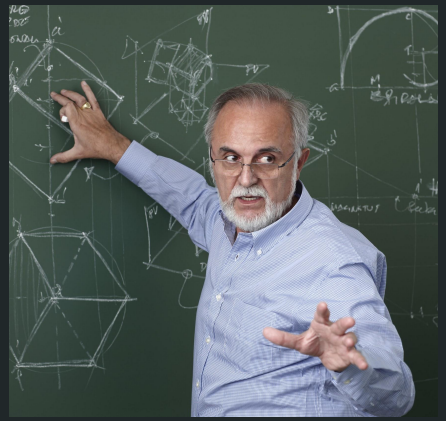


Venues Ground-Truth

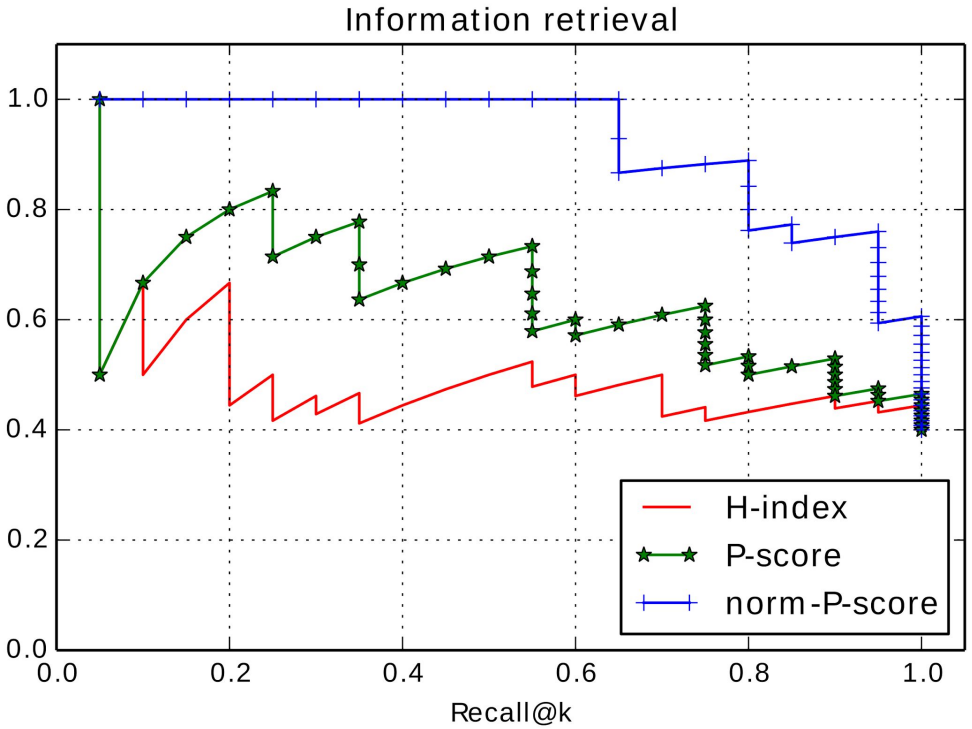


Experimental Results

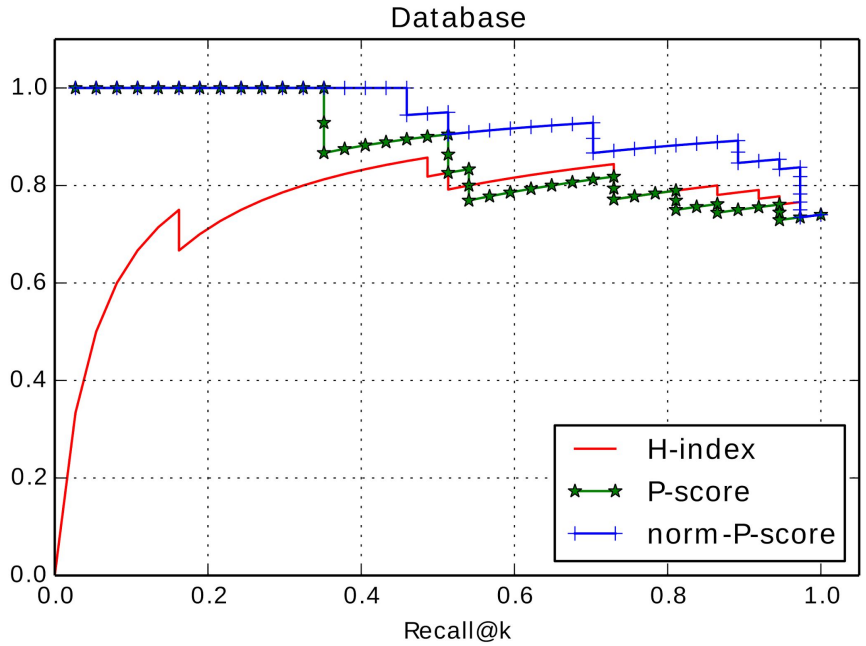
Ranking publication venues and graduate programs in BR and US

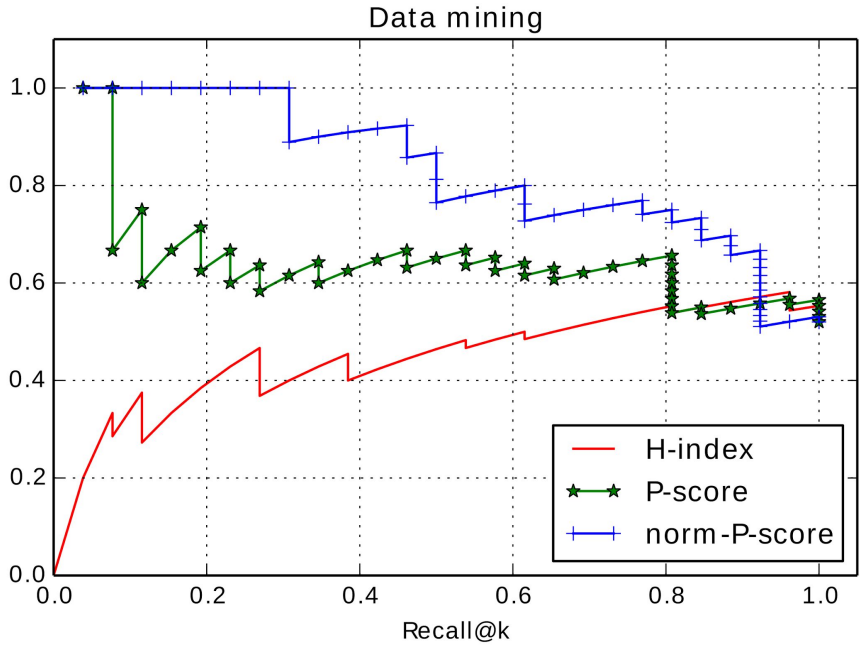


Identifying Venues of a Subarea



Identifying Venues of a Subarea





Ranking of Venues (IR)

#	Standard P-score
1	SIGIR (c)
2	CIKM
3	TREC
4	ECIR
5	CLEF
6	WWW
7	JASIS
8	IPM
9	SIGIR (j)
10	MM
11	JCDL
12	TOIS
13	IR
14	WSDM
15	NTCIR
16	KDD
17	TKDE
18	ACL
19	ICDM
20	SPIRE

norm-P-score
SIGIR (c) CIKM ICTIR IIIX TREC SPIRE ECIR ADCS RIAO IR AIRS NTCIR INEX WSDM TOIS JCDL SIGIR (j) TWEB CLEF LA-WEB

#	Final Ranking
1	SIGIR (c)
2	CIKM
3	TREC
4	ECIR
5	CLEF
6	SIGIR (j)
7	JCDL
8	TOIS
9	IR
10	WSDM
11	NTCIR
12	SPIRE
13	AIRS
14	RIAO
15	INEX
16	IIIX
17	ICTIR
18	ADCS
19	LA-WEB
20	TWEB

Ranking of Graduate Programs (IR)

#	University	P-score
1	Carnegie Mellon University	1
2	University of Massachusetts Amherst	0.8082
3	University of Illinois at Urbana-Champaign	0.6735
4	University of Southern California	0.4541
5	Georgia Institute of Technology	0.4341
6	Stanford University	0.3493
7	University of Illinois at Chicago	0.3409
8	Cornell University	0.3344
9	University of California-Berkeley	0.3337
10	Purdue University	0.3120

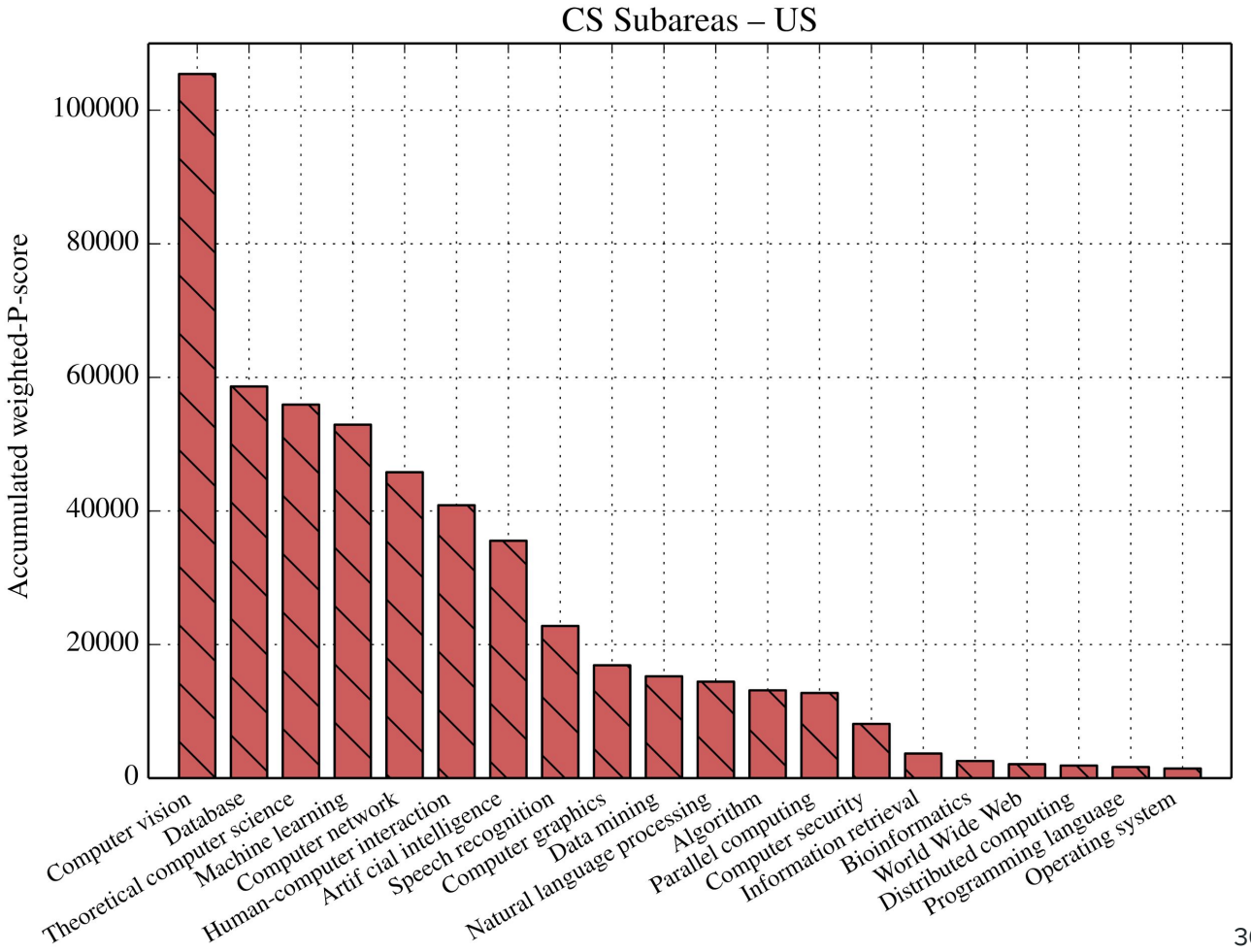
#	University	weighted-P-score
1	University of Massachusetts Amherst	1
2	University of Illinois at Urbana-Champaign	0.4830
3	Carnegie Mellon University	0.4625
4	University of Delaware	0.2452
5	Purdue University	0.2276
6	Northeastern University	0.1633
7	Lehigh University	0.0964
8	Cornell University	0.0552
9	University of Iowa	0.0494
10	University of Illinois at Chicago	0.0477

Ranking of Graduate Programs (IR)

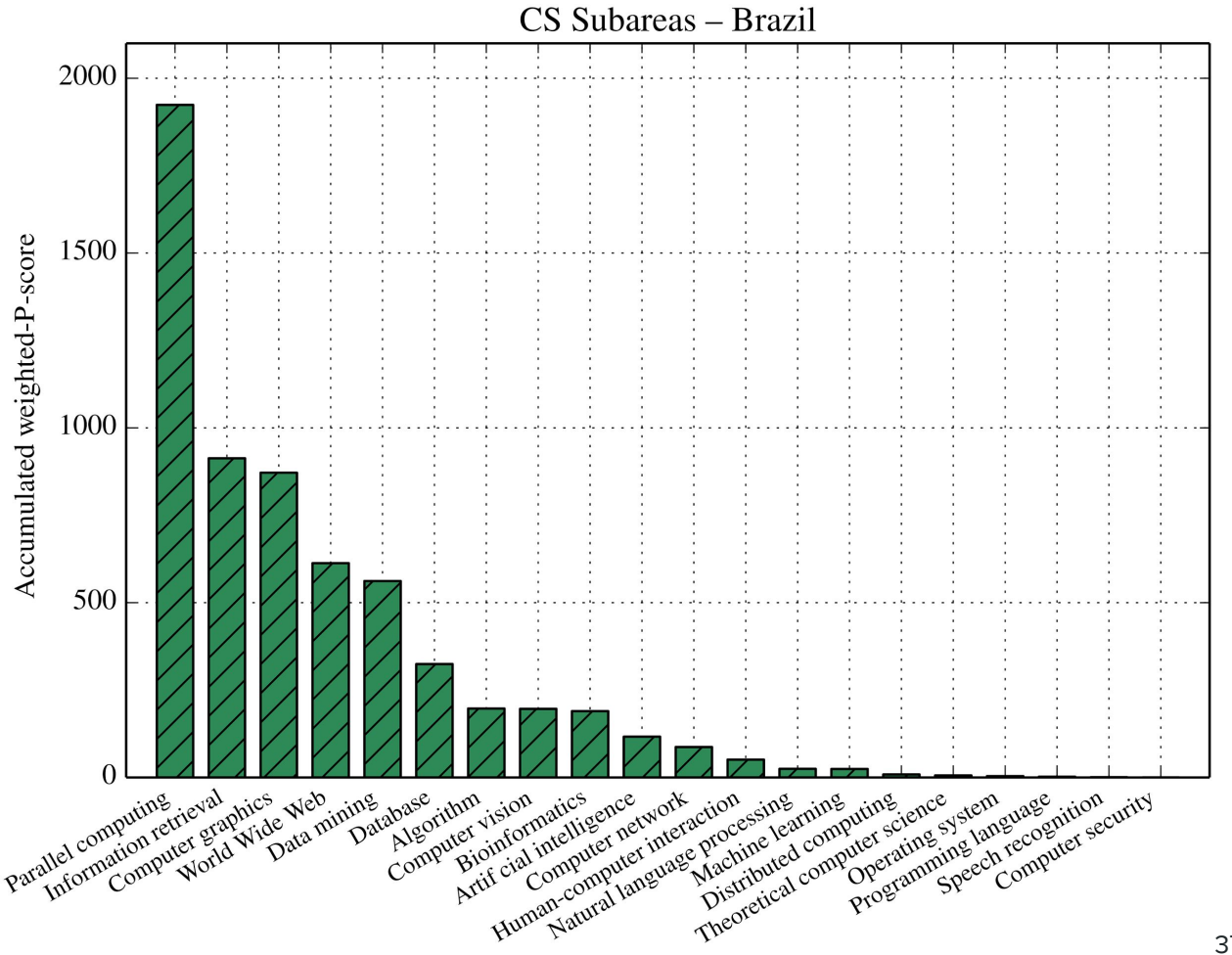
#	Authors' universities
1	University of Massachusetts Amherst #1
2	University of Massachusetts Amherst #2
3	Carnegie Mellon University #1
4	University of Illinois at Urbana-Champaign #1
5	Purdue University
6	University of Delaware
7	Northeastern University
8	University of Illinois at Urbana-Champaign #2
9	Lehigh University
10	Carnegie Mellon University #2
11	University of Iowa
12	University of Illinois at Chicago
13	Georgia Institute of Technology
14	University of Virginia
15	Carnegie Mellon University #3
16	Texas A&M University
17	Cornell University
18	University of Michigan
19	University of Massachusetts Amherst #3
20	New York University

#	University	weighted-P-score
1	University of Massachusetts Amherst	1
2	University of Illinois at Urbana-Champaign	0.4830
3	Carnegie Mellon University	0.4625
4	University of Delaware	0.2452
5	Purdue University	0.2276
6	Northeastern University	0.1633
7	Lehigh University	0.0964
8	Cornell University	0.0552
9	University of Iowa	0.0494
10	University of Illinois at Chicago	0.0477

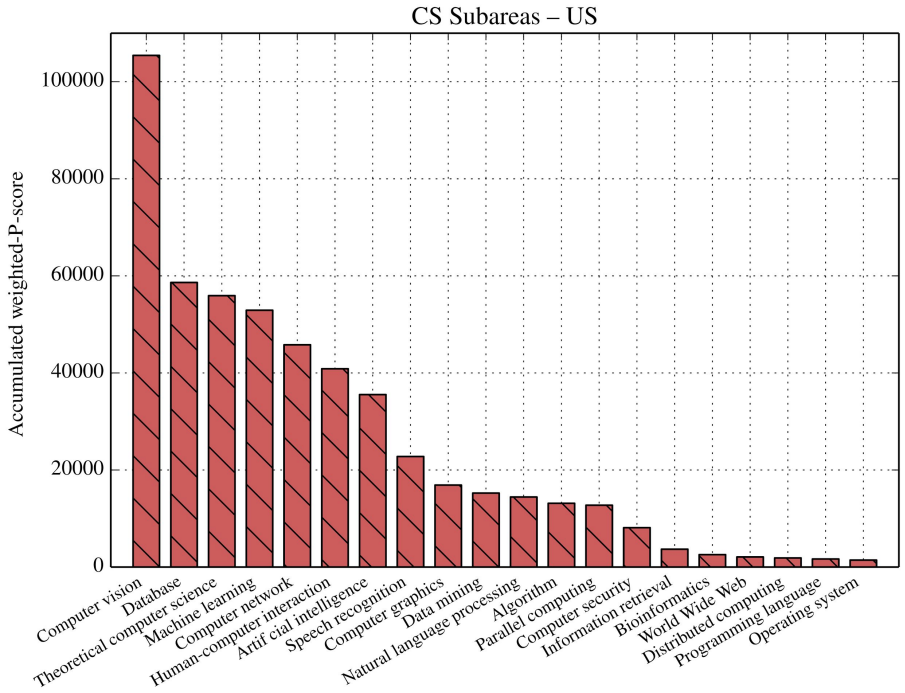
US Research Directions

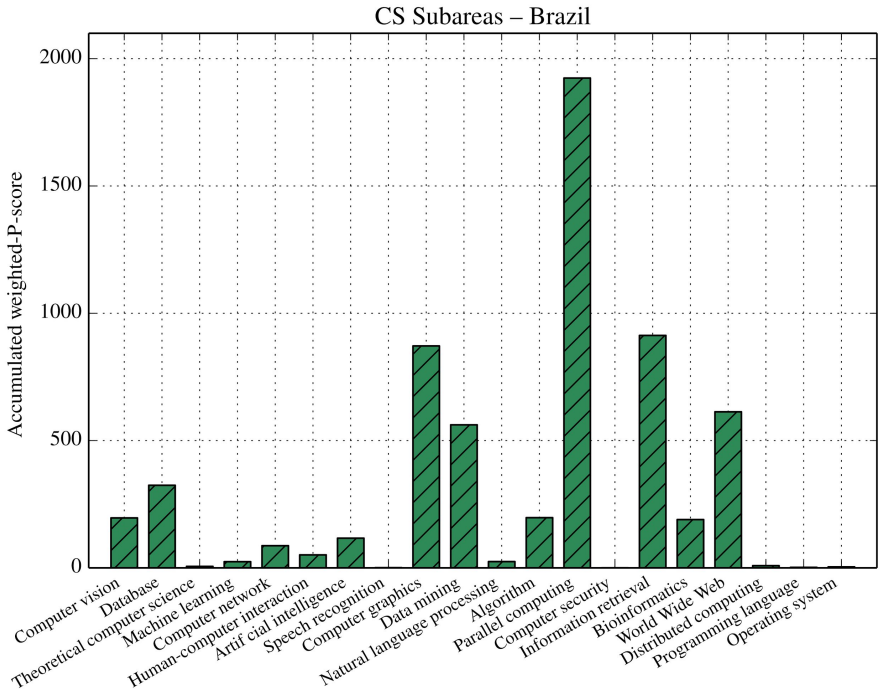


Brazil Research Directions

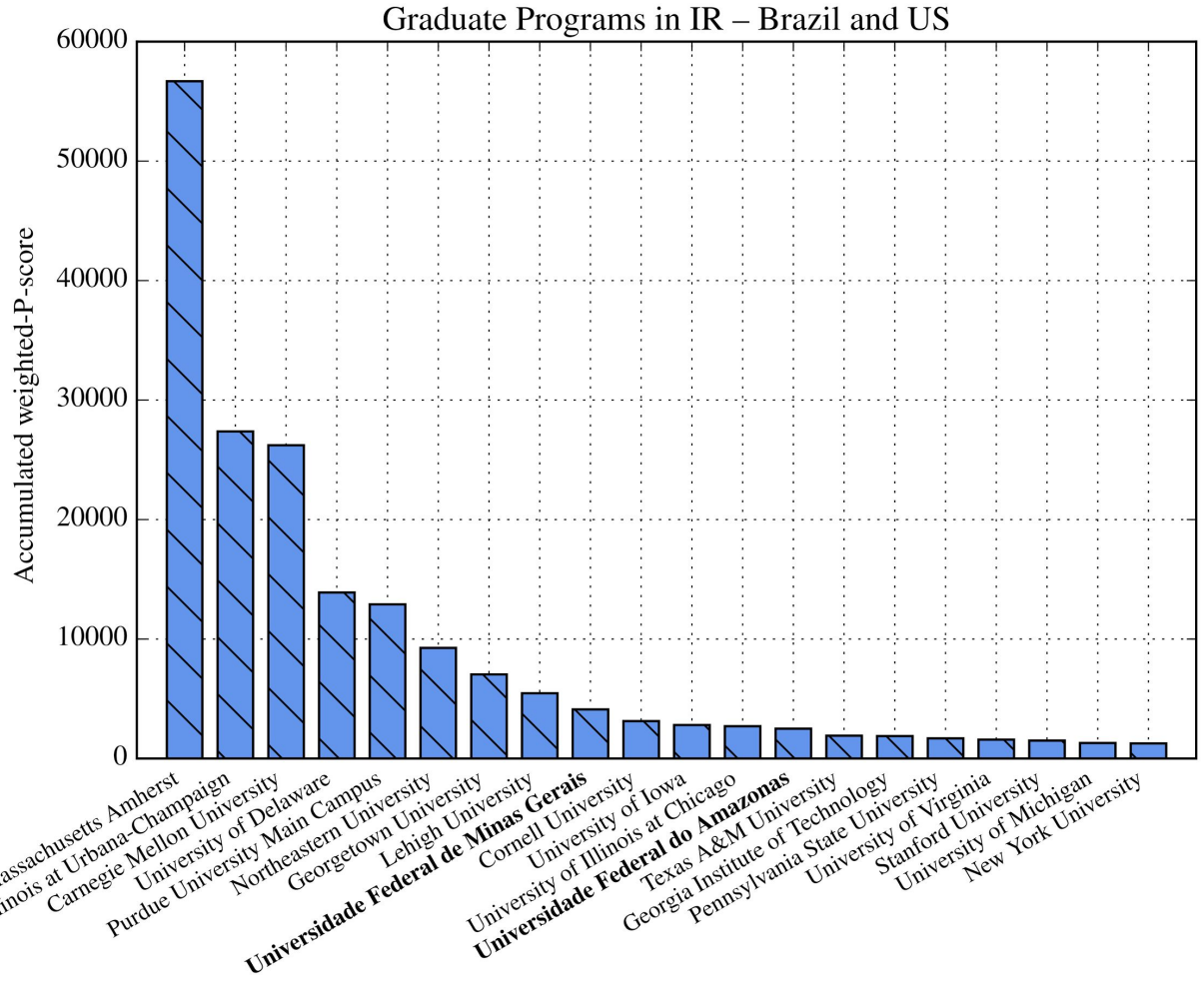


Brazil and US Research Directions

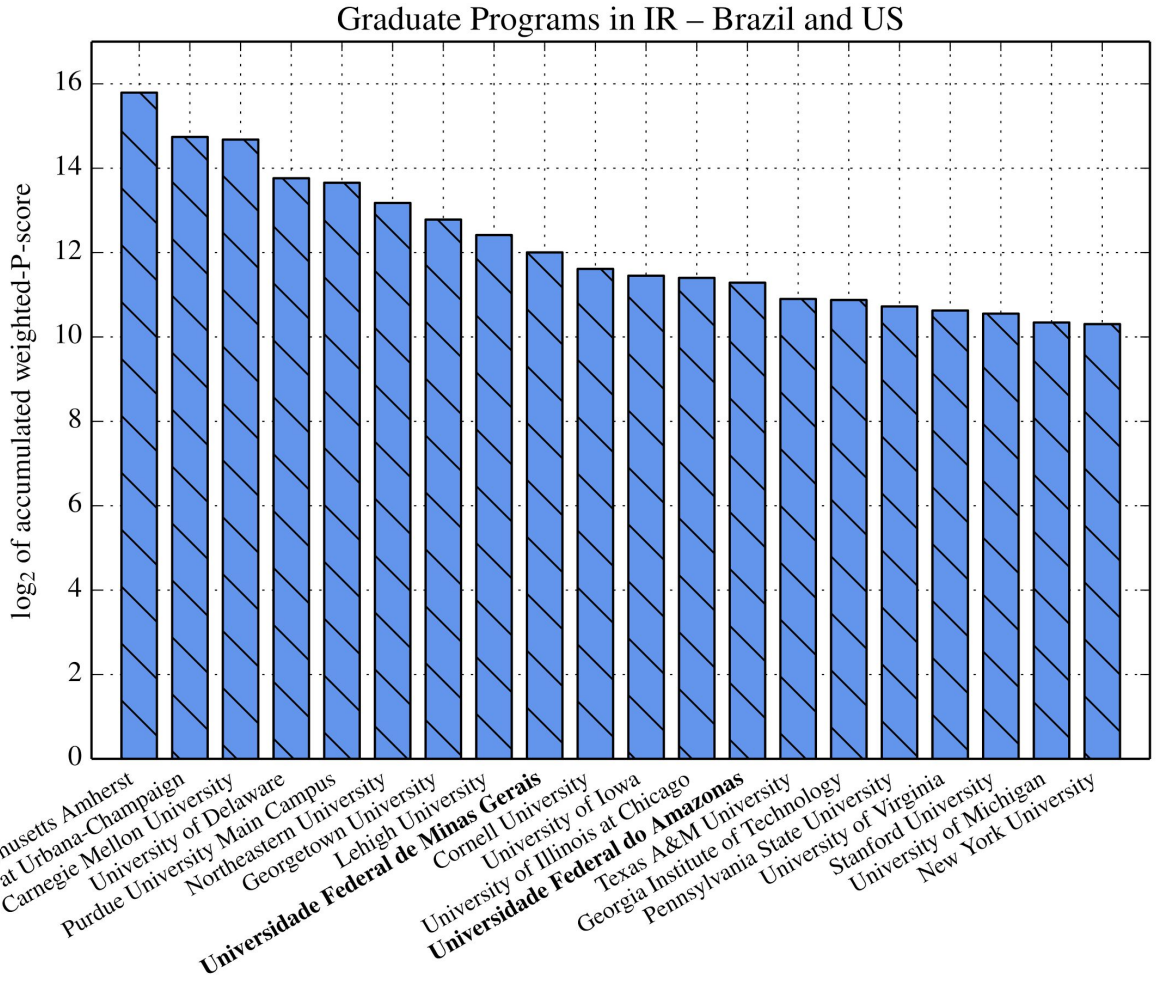




IR Top Graduate Programs

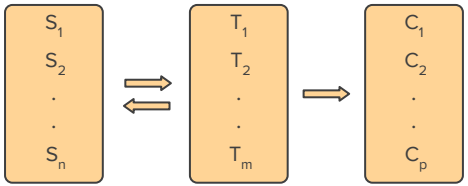


IR Top Graduate Programs



Reputation in CS on a per Subarea Basis

How to **quantify** it?



$$\gamma(a,s) = \frac{1}{|\theta(a)|} \sum_{p \in \theta(a)} \begin{cases} \frac{1}{\log_2(y(0)-y(p)+2)} & \text{if } p \in V_s \\ 0 & \text{otherwise} \end{cases}$$

$$weighted\text{-}P\text{-}score(g,s) = \sum_{a \in \phi(g)} \gamma(a,s) \times \sum_{p \in \theta(a)} \frac{P\text{-}score(venue(p))}{number\_of\_authors(p)}$$

How does it **vary** per CS subarea?

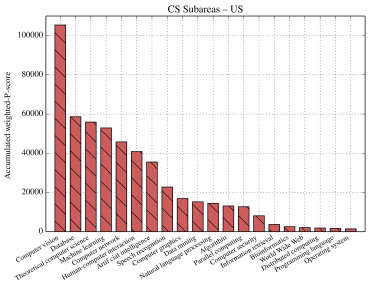
#	Standard P-score
1	SIGIR (c)
2	CIKM
3	TREC
4	ECIR
5	CLEF
6	WWW
7	JASIS
8	IPM
9	SIGIR (j)
10	MM
11	JCDL
12	TOIS
13	IR
14	WSDM
15	NTCIR
16	KDD
17	TKDE
18	ACL
19	ICDM
20	SPIRE

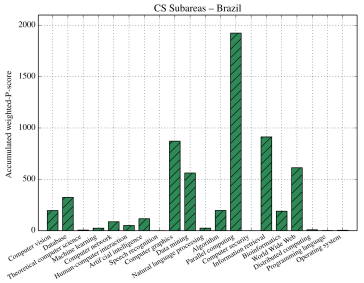
norm-P-score		
SIGIR (c)	CIKM	
ICTIR	IIIX	
TREC	SPIRE	
ECIR	ADCS	
RIAO	IR	AIRS
NTCIR	INEX	
WSDM	TOIS	
JCDL	SIGIR (j)	
TWEB	CLEF	
	LA-WEB	

#	Final Ranking
1	SIGIR (c)
2	CIKM
3	TREC
4	ECIR
5	CLEF
6	SIGIR (j)
7	JCDL
8	TOIS
9	IR
10	WSDM
11	NTCIR
12	SPIRE
13	AIRS
14	RIAO
15	INEX
16	IIIX
17	ICTIR
18	ADCS
19	LA-WEB
20	TWEB

#	University	weighted-P-score
1	University of Massachusetts Amherst	1
2	University of Illinois at Urbana-Champaign	0.4830
3	Carnegie Mellon University	0.4625
4	University of Delaware	0.2452
5	Purdue University	0.2276
6	Northeastern University	0.1633
7	Lehigh University	0.0964
8	Cornell University	0.0552
9	University of Iowa	0.0494
10	University of Illinois at Chicago	0.0477

How do the research in CS in BR and US **differ**?






R

R R R R R R R R R

Conclusions

Solving the venues **encroachment problem** allows us to improve the ranking of graduate programs in a given subarea.

P-score can be used to find core venues of a subarea and also to rank graduate programs on a per subarea basis.

The CS subareas in which **Brazil and US** have major scientific impact are basically disjoint.

What to do next?

Include academic data of other regions of world (e.g. Europe and Asia), identify research trends, perform a temporal analysis of CS communities, validate the reputation model in other broad areas such as Economics.

Reputation in Computer Science on a per Subarea Basis

Alberto Hideki Ueda
Berthier Ribeiro-Neto and Nivio Ziviani

July 14, 2017

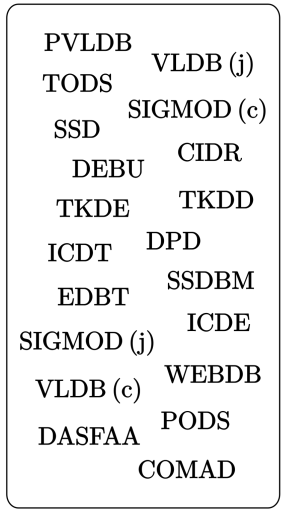
ueda@dcc.ufmg.br

“We have a responsibility to better inform the public, by providing them with relevant information.”

Moshe Y. Vardi, editor-in-chief of Communications of the ACM 2016, about academic rankings

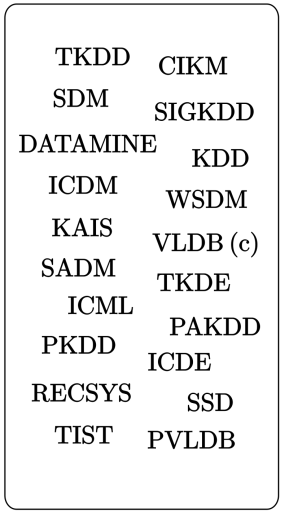
Ranking of Venues (DB and DM)

#	Standard P-score
1	SIGMOD (c)
2	ICDE
3	VLDB (c)
4	PVLDB
5	TKDE
6	DEBU
7	SIGMOD (j)
8	EDBT
9	CIKM
10	PODS
11	TODS
12	KDD
13	VLDB (j)
14	WWW
15	ICDM
16	DASFAA
17	SSDBM
18	IS
19	ICDT
20	CIDR

norm-P-score


#	Final Ranking
1	SIGMOD (c)
2	ICDE
3	VLDB (c)
4	PVLDB
5	TKDE
6	DEBU
7	SIGMOD (j)
8	EDBT
9	PODS
10	TODS
11	VLDB (j)
12	DASFAA
13	SSDBM
14	ICDT
15	CIDR
16	WEBDB
17	SSD
18	DPD
19	COMAD
20	TKDD

#	Standard P-score
1	KDD
2	ICDM
3	CIKM
4	ICDE
5	ICML
6	SDM
7	TKDE
8	WWW
9	SIGMOD
10	AAAI
11	PKDD
12	NIPS
13	PAKDD
14	VLDB (c)
15	SIGIR
16	SIGKDD
17	DATAMINE
18	KAIS
19	JMLR
20	PVLDB

norm-P-score


#	Final Ranking
1	KDD
2	ICDM
3	CIKM
4	ICDE
5	ICML
6	SDM
7	TKDE
8	PKDD
9	PAKDD
10	SIGKDD
11	DATAMINE
12	KAIS
13	PVLDB
14	WSDM
15	TKDD
16	VLDB (c)
17	RECSYS
18	TIST
19	SSD
20	SADM

Ranking of Grad. Programs (DB and DM)

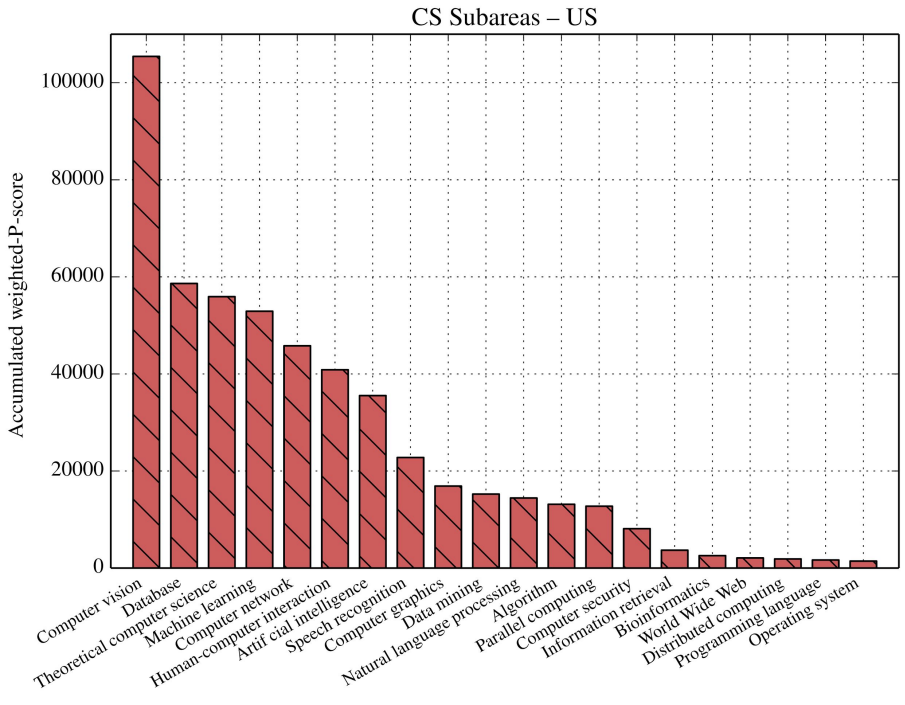
#	University	weighted-P-score
1	University of Wisconsin-Madison	1
2	Stanford University	0.6570
3	University of Illinois at Urbana-Champaign	0.5687
4	Massachusetts Institute of Technology	0.4975
5	Duke University	0.4616
6	University of Massachusetts Amherst	0.4243
7	University of Michigan	0.4195
8	University of California-Irvine	0.4120
9	University of Maryland-College Park	0.4101
10	University of California-Santa Cruz	0.3982

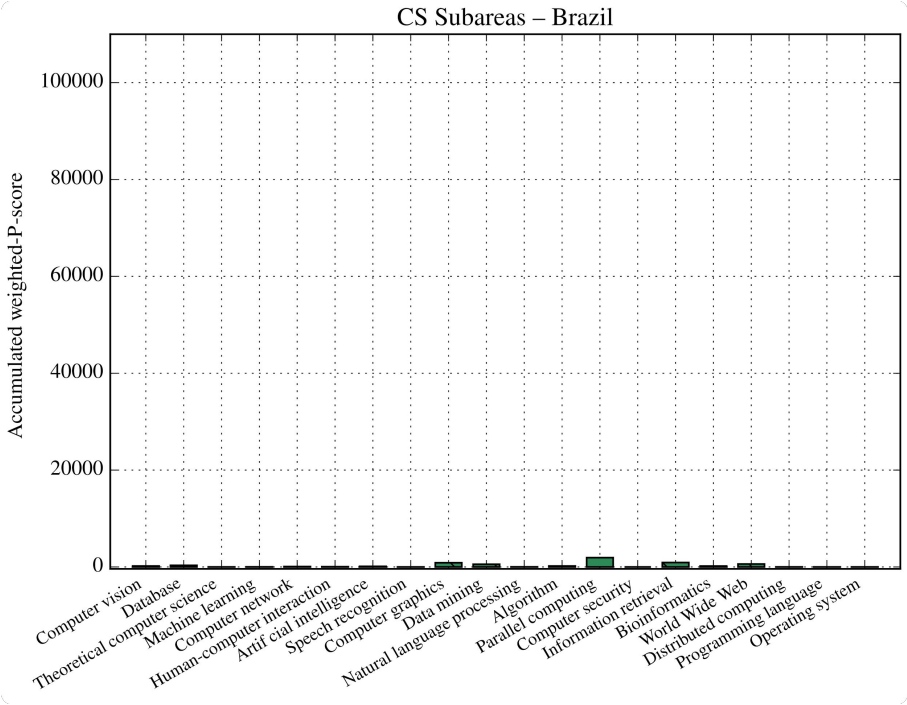
#	University	weighted-P-score
1	University of Illinois at Chicago	1
2	Carnegie Mellon University	0.6857
3	University of Illinois at Urbana-Champaign	0.6344
4	University of Minnesota	0.5350
5	Arizona State University	0.4276
6	University of California-Riverside	0.4212
7	Georgia Institute of Technology	0.3955
8	University of Michigan	0.3275
9	Rensselaer Polytechnic Institute	0.2761
10	University of California-Davis	0.2593

Standard P-score Suggested Configurations

	Venues	Authors	Groups
Groups	$G^* \rightleftharpoons V \rightarrow V$	$G^* \rightleftharpoons V \rightarrow A$	$G^* \rightleftharpoons V \rightarrow G$
Authors	$A^* \rightleftharpoons V \rightarrow V$	$A^* \rightleftharpoons V \rightarrow A$	$A^* \rightleftharpoons V \rightarrow G$
Venues	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow V$	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow A$	$V^* \rightleftharpoons A \therefore A^* \rightleftharpoons V \rightarrow G$

Brazil and US Research Directions





“When a measure becomes a target, it ceases to be a good measure.”

Goodhart's Law,
Charles Goodhart, Economist, 1975