Indicators of scientific and technological culture: Wikipedia

Carlos G. Figuerola, Angel Zazo Rodríguez, José L. Alonso Berrocal

Universidad de Salamanca. Instituto ECyT
(*figue—zazo—berrocal@usal.es)*
http://reina.usal.es

Introduction

WIKIPEDIA
La enciclopedia libre

- milions of users every day
- built in a collaborative way
- good to social perception of Sci & Tech
- no matter about their quality, our aim is to analyze how people see Sci & Tech

Introduction

- downloads of the whole Wikipedia available
  - organized by languages
  - database includes administrative data
  - database structure is very complex
- we work with Spanish Wikipedia of January 2012

Introduction

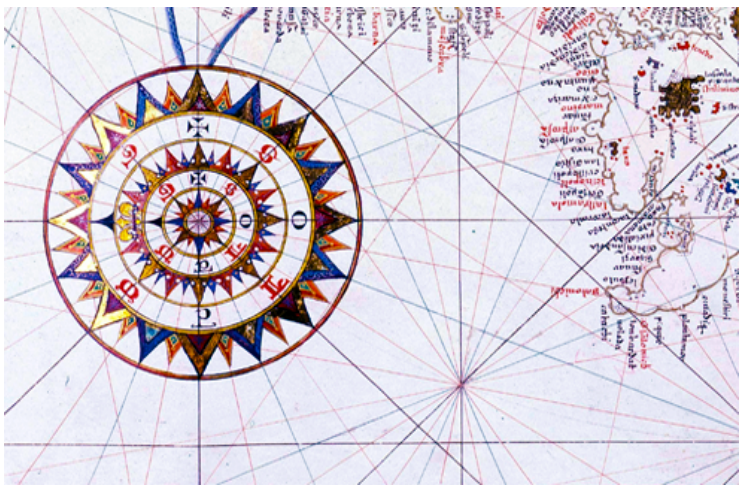
Some data (Spanish Wikipedia Jan 2012):

After cleaning:

- 865,393 articles
- 21,543,603 internal links
- 63,669 categories

Selecting articles on Sci & Tech

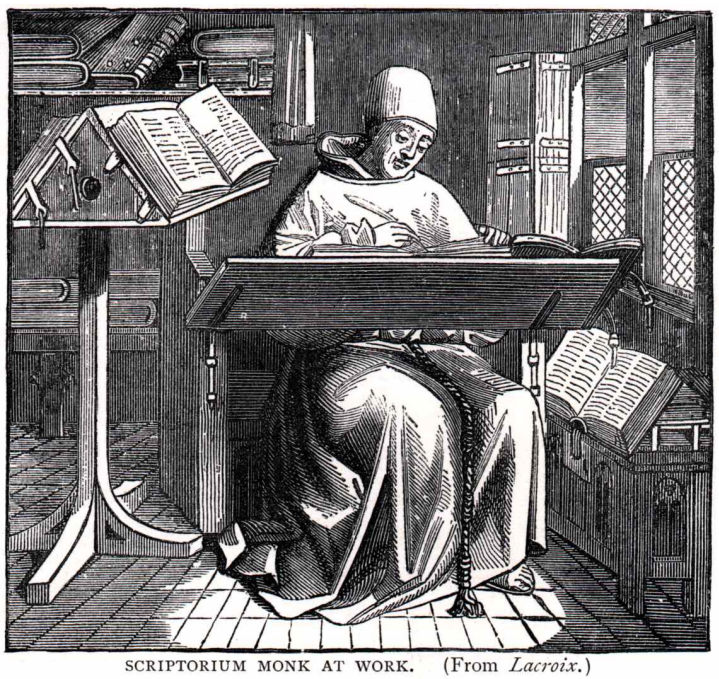
865 K articles are too much!



- we can't select articles by hand
- sampling discarded (looks sloppy in this context)
- fortunately, articles come inside categories
- categories's number is lower

Selecting articles on Sci & Tech

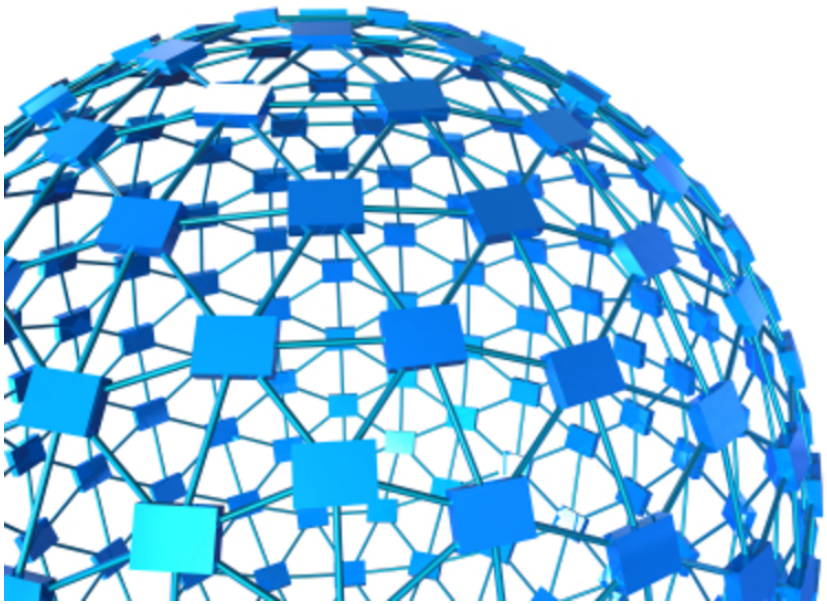
Categories work as content tags in Wikipedia



SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

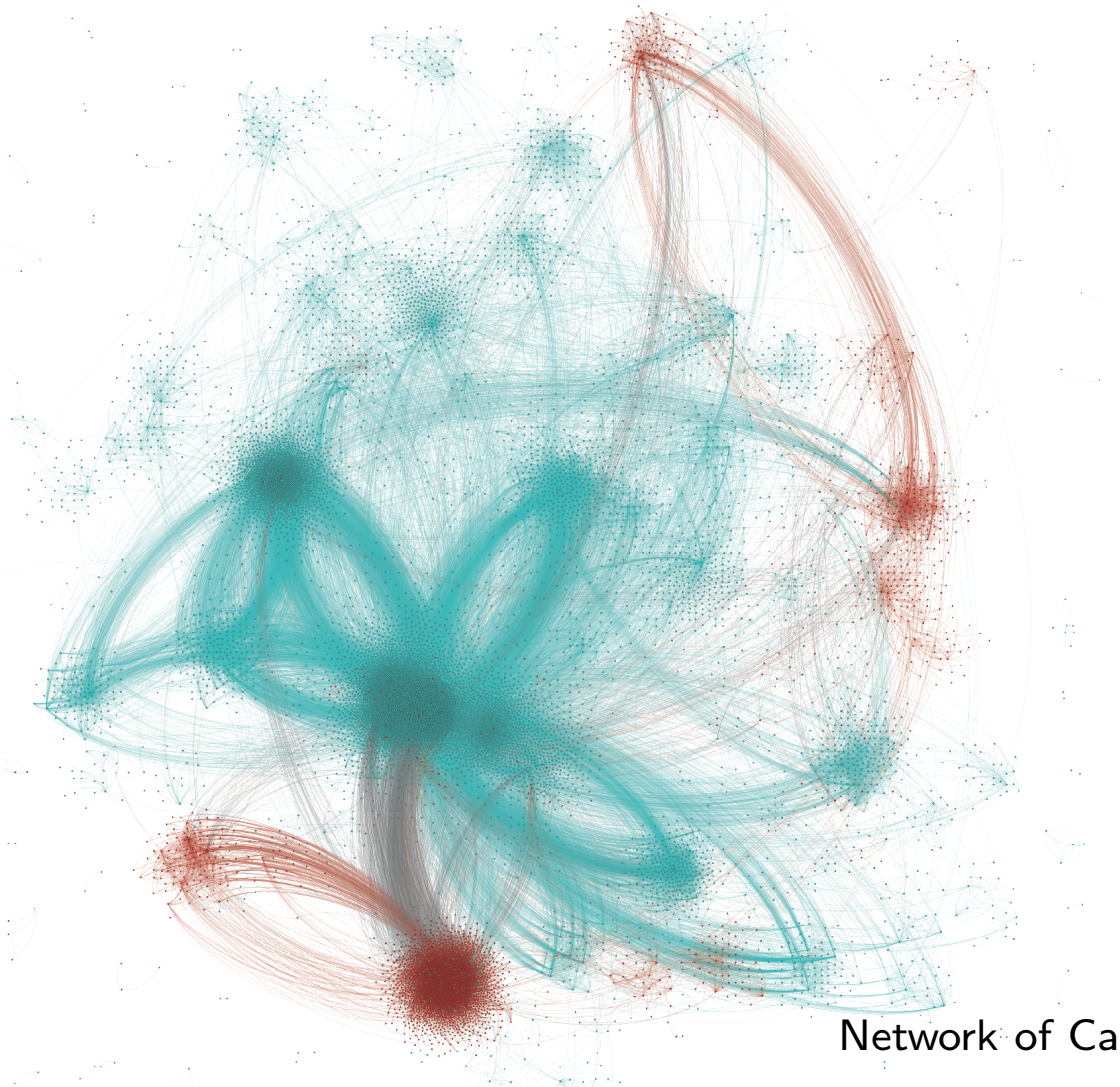
- an author can assign multiple categories to an article
- an author can create a new category at any time
- as a result, we have many categories with very few articles

Selecting articles on Sci & Tech



- we collected articles in each category
- we grouped hyperlinks between articles from every category to each other
- we built a network with categories as nodes and grouped links as edges

Selecting articles on Sci & Tech

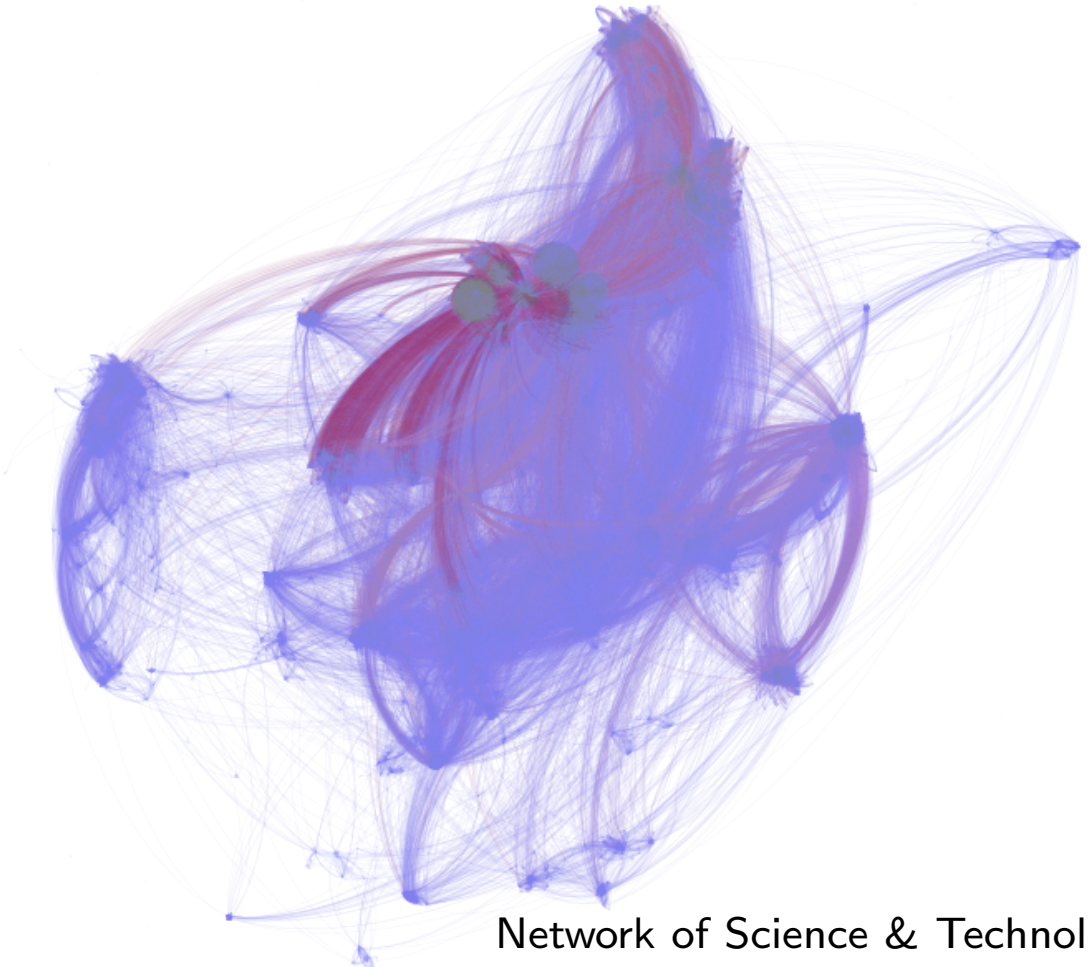


Network of Categories

Selecting articles on Sci & Tech

- we can to apply techiques from Social Networks Analysis (SNA)
- Communities Detection:
  - InfoMap algorithm (it takes in account weight and direction of edges)
  - it produces 839 comunities, they can be evaluated by hand
  - 116 communities were classified as Sci & Tech
  - these communities have 3,471 categories
  - which, in their turn, have **94,797** different articles

Selecting articles on Sci & Tech



Network of Science & Technology articles

Selecting articles on Sci & Tech

However ...

- many of articles are a simple templates for taxonomies's elements (for example botanical and zoological species, asteroids ...)

So ...

- we removed articles:
  - having a template for taxons AND
  - having less than a specified number of characters AND
  - having only 1 edition of their content

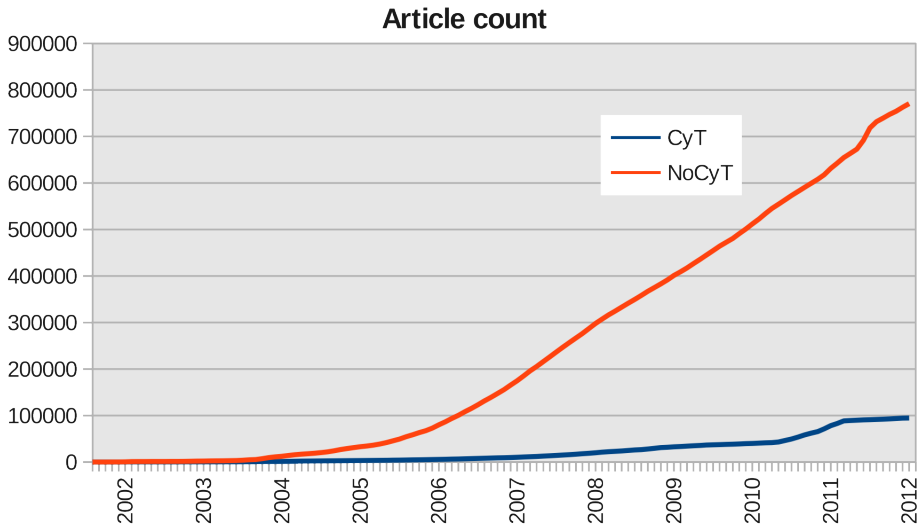
After that, we have **29,639** significant articles about Sci & Tech

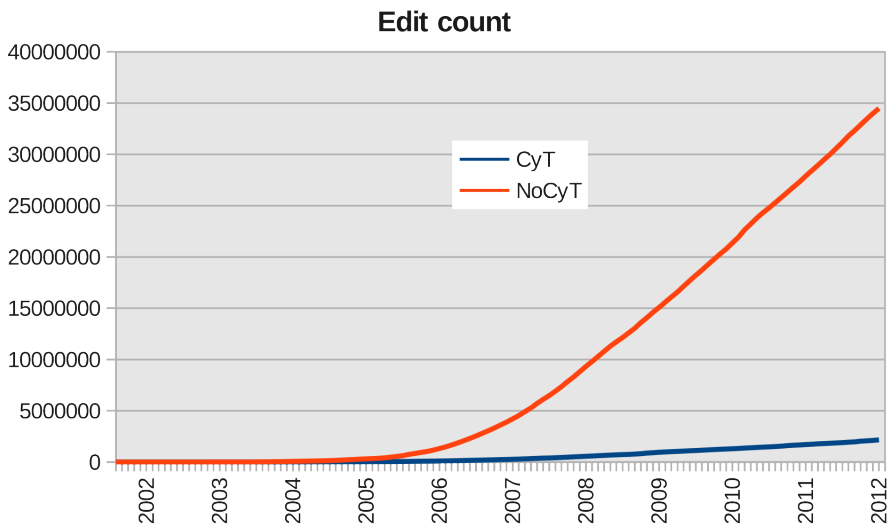
Selecting articles on Sci & Tech

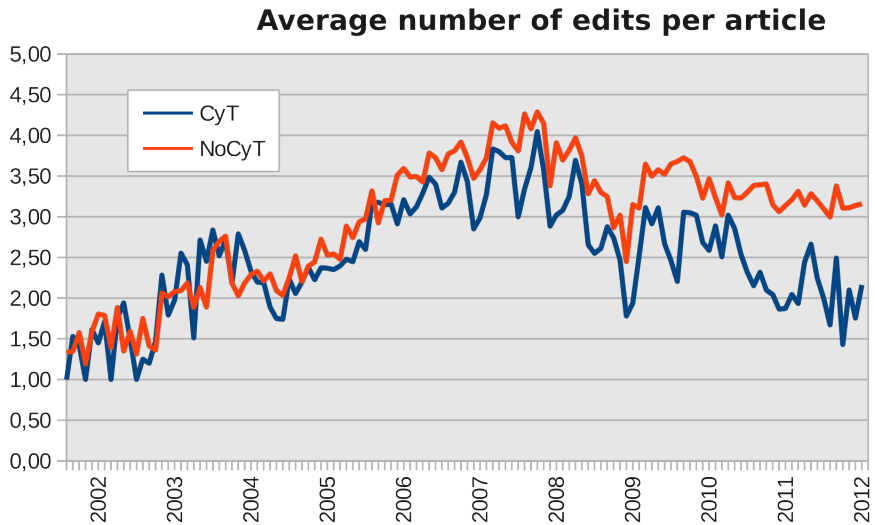
Authoring/Editing in Wikipedia

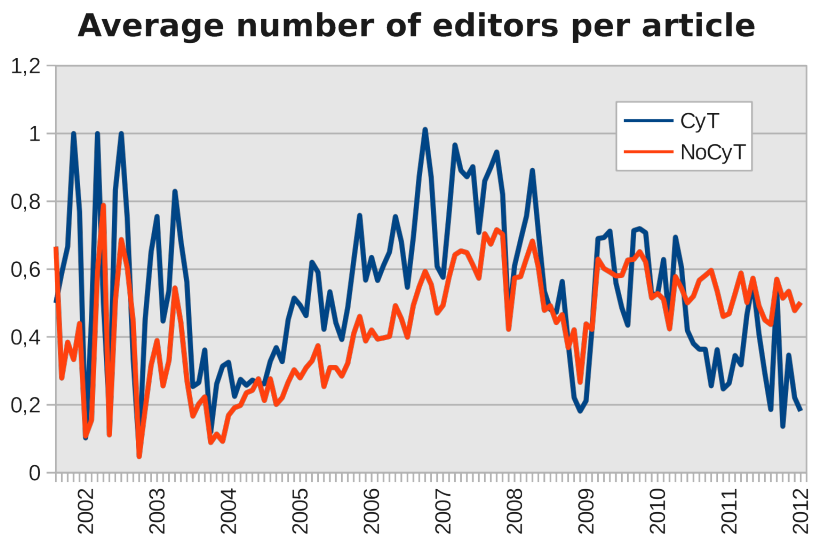
Authoring/editing in Wikipedia

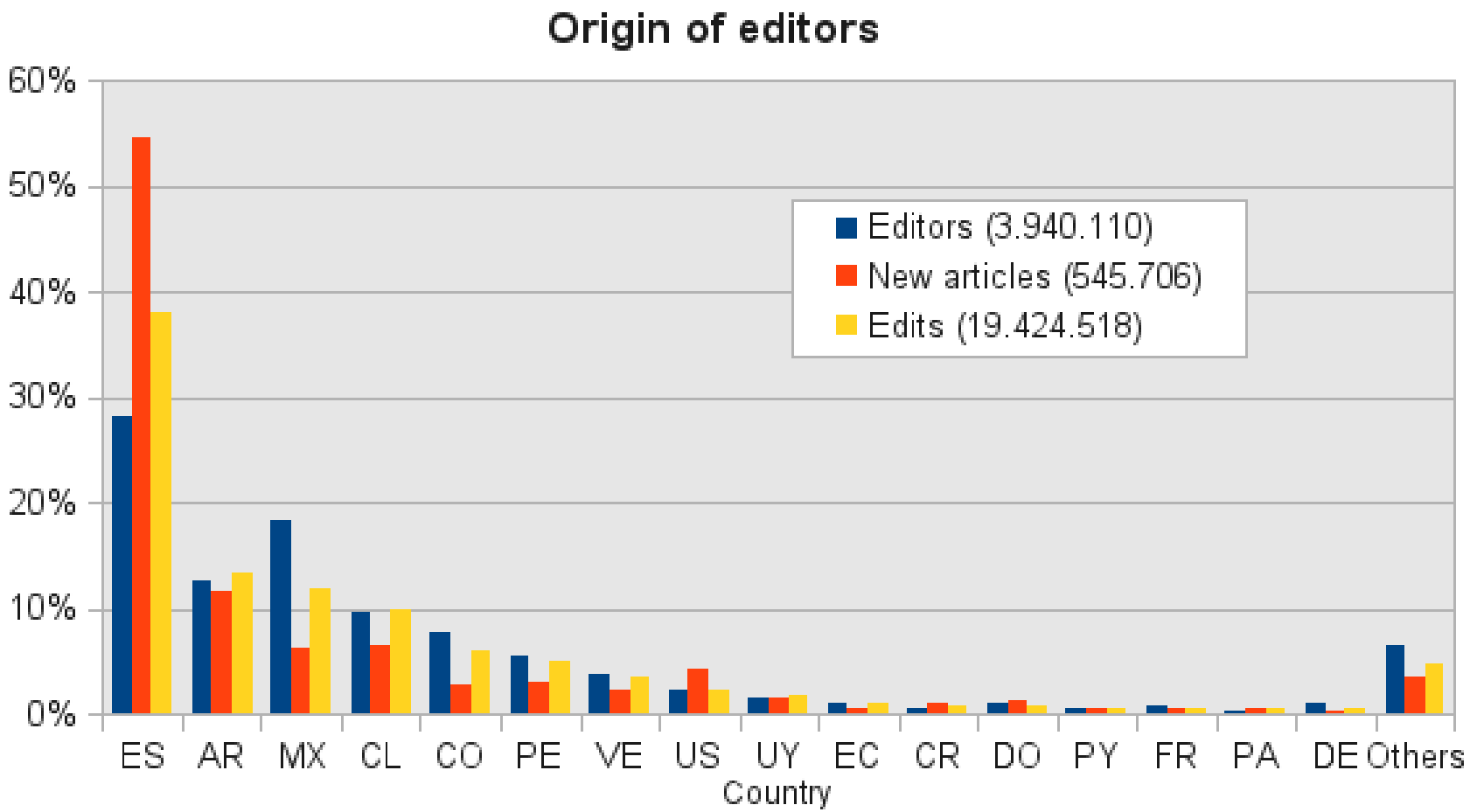
Editors	#	Article's Edits		
		CyT	No CyT	All
Registered editors (no admin)	345,481	468,692 (4.4%)	10,288,585	10,757,277
Administrative editors (sysop, bureaucrat, rollbacker, checkuser...)	621	496,596 (7.1%)	6,521,165	7,017,761
Bots	334	704,486 (9.4%)	6,780,667	7,485,153
Unregistered editors	3,934,686	497,571 (4.4%)	10,898,498	11,396,069
Total	4,281,122	2,167,345 (5.9%)	34,488,915	36,656,260

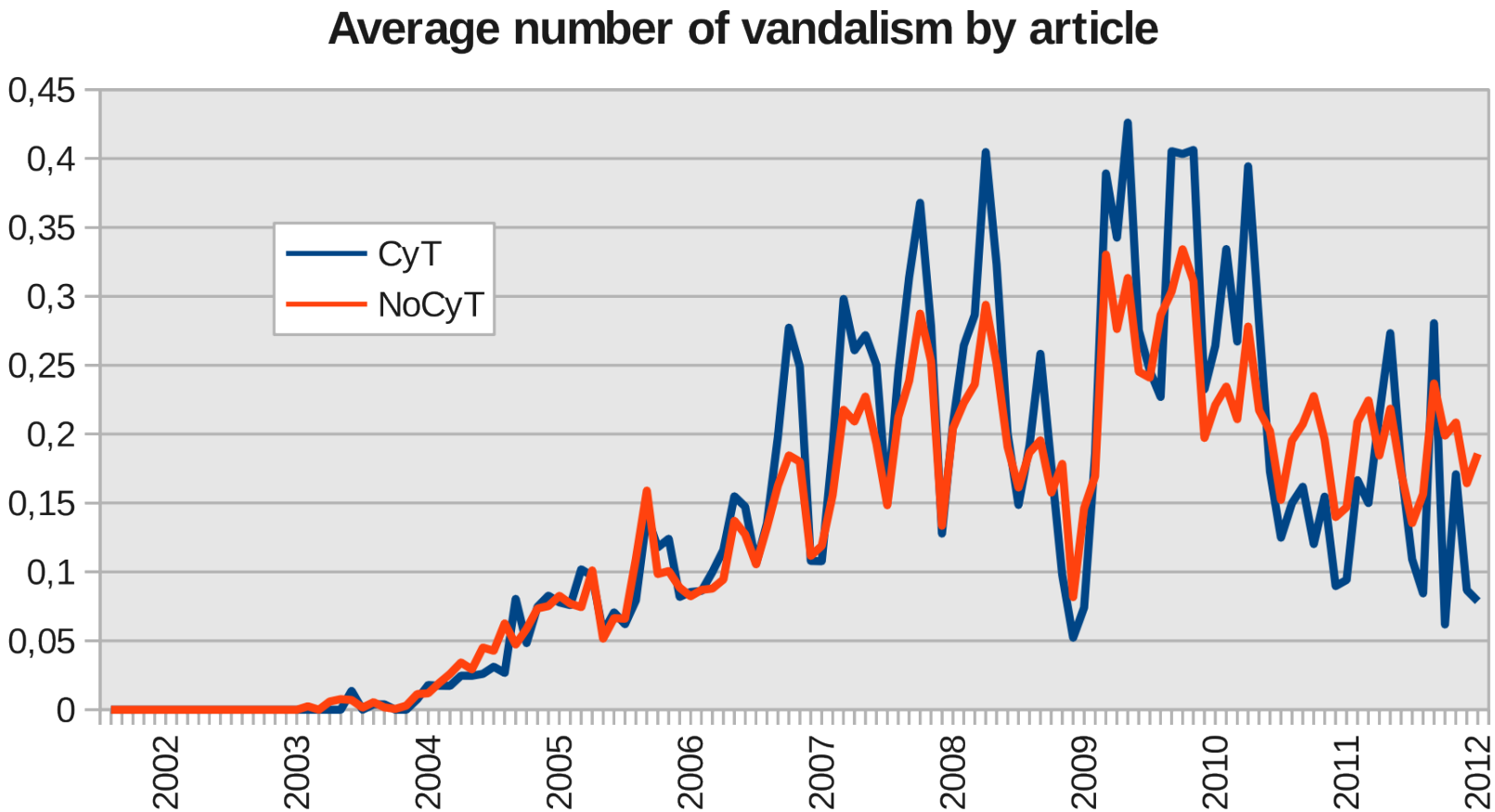

Article count
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
CyT
NoCyT
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012


Edit count
40000000
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
CyT
NoCyT
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012


Average number of edits per article
5,00
4,50
4,00
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
CyT
NoCyT
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012


Average number of editors per article
1,2
1
0,8
0,6
0,4
0,2
0
CyT
NoCyT
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012


Origin of editors
60%
50%
40%
30%
20%
10%
0%
Editors (3.940.110)
New articles (545.706)
Edits (19.424.518)
ES
AR
MX
CL
CO
PE
VE
US
UY
EC
CR
DO
PY
FR
PA
DE
Others
Country


Average number of vandalism by article
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
CyT
NoCyT
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Authoring/editing in Wikipedia

	S & T	no S & T
Creation		
Created by	Admins	Admin/regist.
Edition		
Editors per art.	13.3	22.2
Edits per art.	22.9	42.4
Vandalism		
Arts. vandalized	12 %	25%

Science & Tech articles seem to be less participatory

Applying SNA techniques

Applying SNA Techniques

How link Sci & Tech articles with the others articles?

- articles of Sci & Tech produce 1,015,129 links out
- of them, 428,366 point to no Sci & Tech arts.
- most of linked no Sci & Tech arts. are geographic places (countries, cities, etc) or years of dates

Applying SNA Techniques

linked art.	No. of links		
Wikimedia_Commons	8,202	Industria	101
ISBN	2,204	Televisión	97
Creative_Commons	1,339	Vinagre	96
Latín	923	Universo	93
Idioma_griego	778	Café	87
Nombre_común	771	Roseta	87
Idioma_árabe	280	Unesco	87
M	239	Orden	87
Idioma_francés	188	Plinio_el_Viejo	84
Clima	173	Ganado	83
Cultivar	159	Jardín	80
Carne	139	Ensalada	75
Alimento	139	Infusión	75
Facsímil	117	Cristianismo	75
Aceite	116	Gastronomía	74
Chocolate	115	Cerámica	72
Queso	106	Sopa	72
Vino	105	Patrimonio_de_la_Humanidad	69
Primera_Guerra_Mundial	102		

Table : Non Sci & Tech arts. most linked by Sci & Tech (exluding geographic places and dates)

Kilómetro_cuadrado	Segundo	Plata	Bosque
Área	Desarrollador_de_videojuegos	Madera	Longitud
Altitud	Género_de_videojuegos	Química	Biología
Metro	Metro_cúbico	Hierro	Especie_bajo_preocupa ...
Población	Distribuidora_de_videojuegos	Agua_dulce	Google
Verificabilidad	Dominio_público	Fotografía	Software
Licencia_de_documentación...	Hectárea	Sistema_operativo	Gramo
Sitio_web	Escala_temporal_geol...	Cáncer	ITunes
Kilómetro	Medicina	Tonelada	Frecuencia
Especie	Agua	Oxígeno	Luna
Distancia	Aves	Hoja	Nudo_(unidad)
Kilogramo	Tierra	Cuenca_hidrográfica	Grado_Celsius
Libra_(unidad_de_masa)	Oro	Multimedia	Ingeniería
Agricultura	Médico	Centímetro	NASA
Unión_Internacional_para_la...	Temperatura	Hidrógeno	Licencia_de_software
Estado_de_conservación	Mar	Internet_Archive	Astronomía
Nivel_del_mar	Física	Psicología	Unión_Internacional_de_Quí...
Arquitectura	Flor	Petróleo	Acero
Digital_object_identifier	Endemismo	Ciencia	Online_Computer_Library...
Latitud	Matemáticas	Cobre	Tecnología
Río	Botánica	Microsoft_Windows	Flora
DVD	Mitología_griega	Jardín_botánico	Hábitat
Videojuego	Clima_tropical	Sol	Carbón
Internet	Árbol	Carbono	Neolítico
Plataforma_(informática)	Google_Earth		

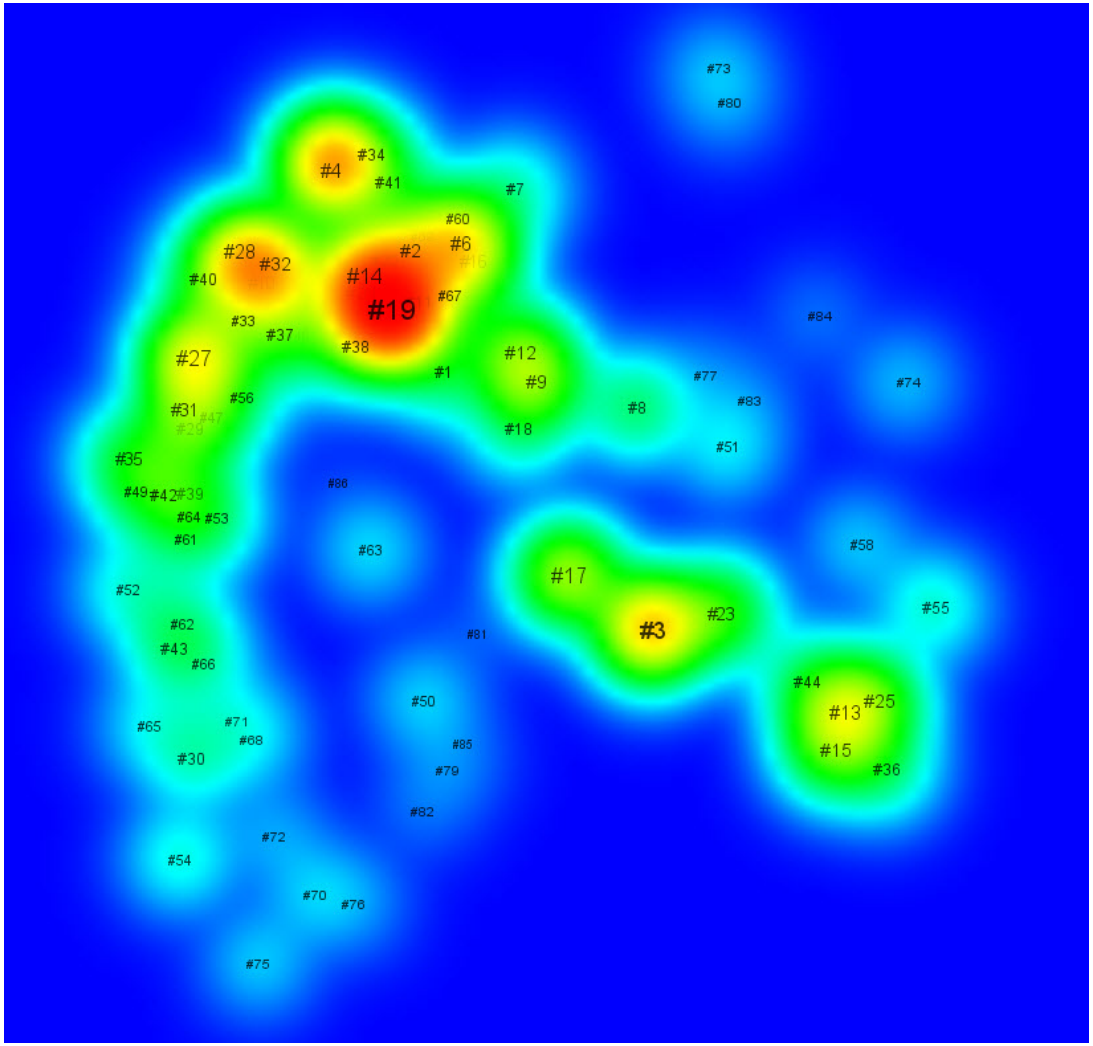
Table : The most linked arts. about Sci & Tech from the rest of Wikipedia

GNU/Linux	Mozilla_Firefox
Internet	Sistema_operativo
Carl_Gustav_Jung	Hardware
Ubuntu	Agua
Amor	Sigmund_Freud
Cannabis_sativa	Videojuego
Software	Windows_Vista
Microprocesador	Sistema_Solar
Microsoft	Ecosistema
Microsoft_Windows	Disco_duro
Electricidad	Informática
Microsoft_Excel	Quiromasaje
Ácido_desoxirribonucleico	Potenciación
Windows_7	Aparato_digestivo
Cáncer	Psicoanálisis
Proteína	Corazón
Windows_XP	Estado_de_agregación_de_la_materia
Calentamiento_global	Energía
Sida	Salud
Desarrollo_sostenible	Web_2,0
PHP	Dirección_IP
Software_libre	Luna
Infecciones_de_transmisión_sexual	Ojo_humano
Equus_ferus_caballus	Homo_sapiens
Animalia	Virus_informático

Table : The most reviewed articles about Sci & Tech

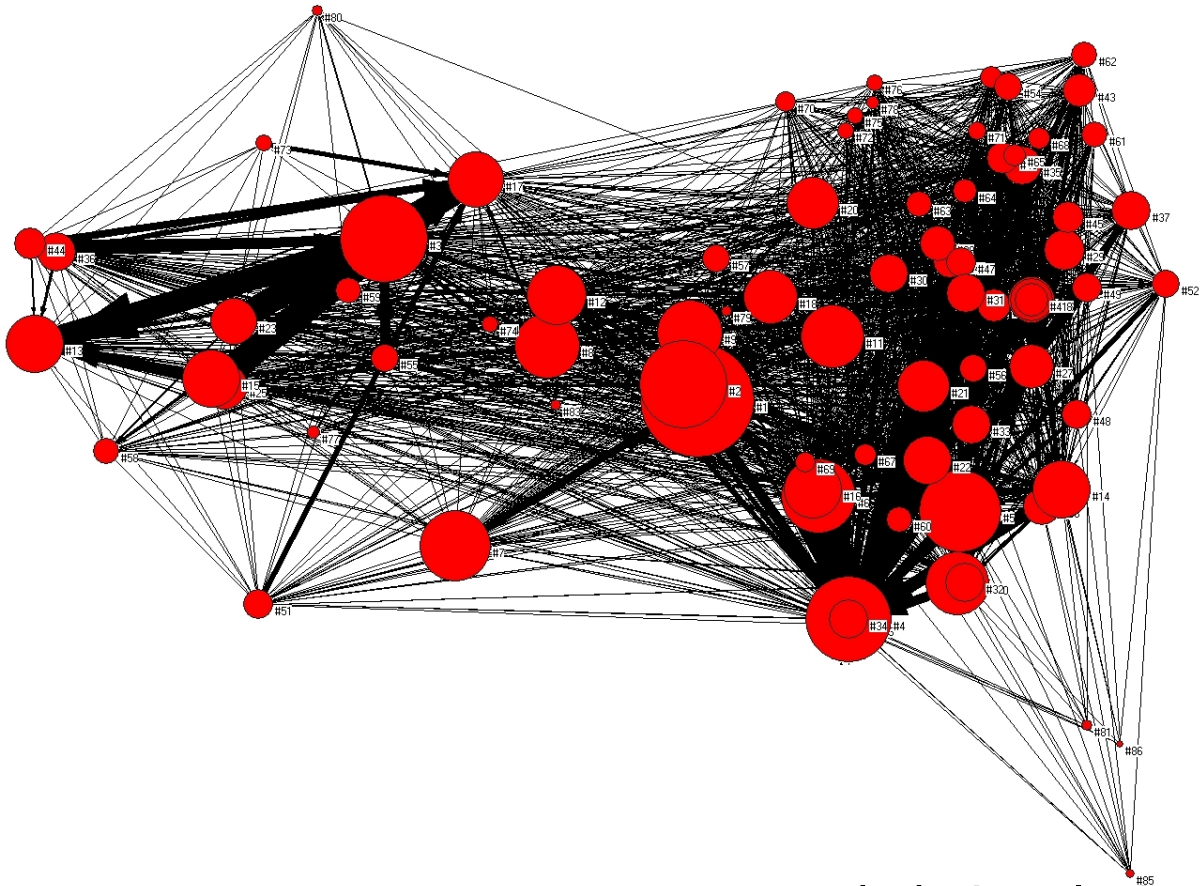
Applying SNA Techniques

Detecting communities ...



Carlos G. Figuerola, Angel Zazo Rodríguez, José L. Alonso Berrocal: Indicators of scientific and technological culture: Wikipedia

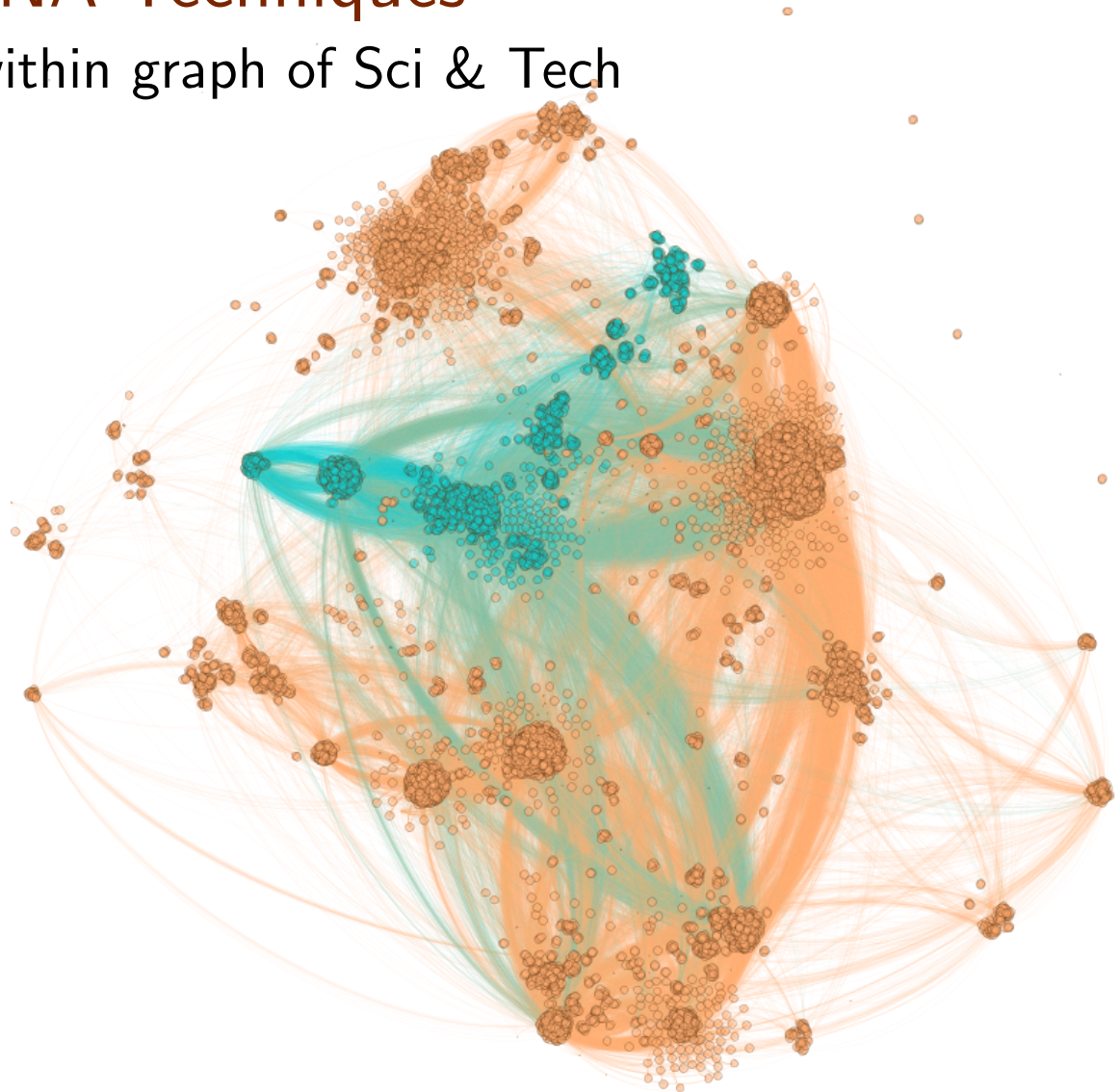
Applying SNA Techniques



... and their relationships.

Applying SNA Techniques

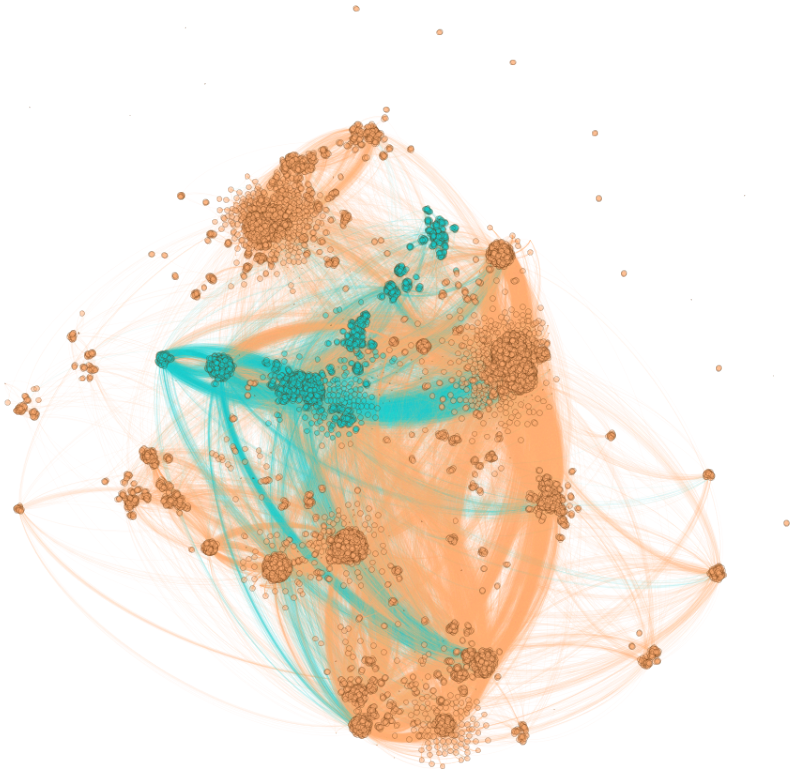
Life Sci within graph of Sci & Tech



Carlos G. Figuerola, Angel Zazo Rodríguez, José L. Alonso Berrocal: Indicators of scientific and technological culture: Wikipedia

Applying SNA Techniques

Links **from** Life Sci.



Links **to** Life Sci.

ISBN	247	Símbolo	10	Cerveza	7
Filosofía	35	Sociología	10	Antiguo_Egipto	7
Anatomía_de_Gray	31	Ganado	10	Culpa	7
Alimento	25	Felicidad	10	Eros	7
Escritor	21	País	10	Homosexualidad	7
Tabaco	18	Queso	9	Derechos_de_los_animales	7
Educación	18	Estilo_de_vida	9	Edad	7
Aprendizaje	18	Literatura	9	Real_Academia_Española	7
Población	18	Mitología	9	Feminismo	7
Religión	17	Conocimiento	9	Paz	7
Familia	17	Economía	9	Concepto	7
Persona	16	Estado	9	Cigarrillo	7
Sociedad	16	Etnia	9	Alquimia	7
Cultura	15	Suicidio	9	Ensueño	7
Ética	14	Escuela	9	Mito	7
Arte	14	Vino	9	Pensamiento	7
Política	14	Justicia	8	Tercer_mundo	6
Derechos_humanos	13	Tiempo	8	Espiritualidad	6
Cristianismo	12	Pobreza	8	Derecho	6
Alimentación	12	Hambre	8	Chocolate	6
Dios	11	Carne	8	Moral	6
Raza	11	Occidente	8	Violencia	6
Meditación	11	Biblia	8	Alma	6
Internet_Archive	10	Romanticismo	8	Creencia	6
Budismo	10	Licor	7	Judaísmo	6

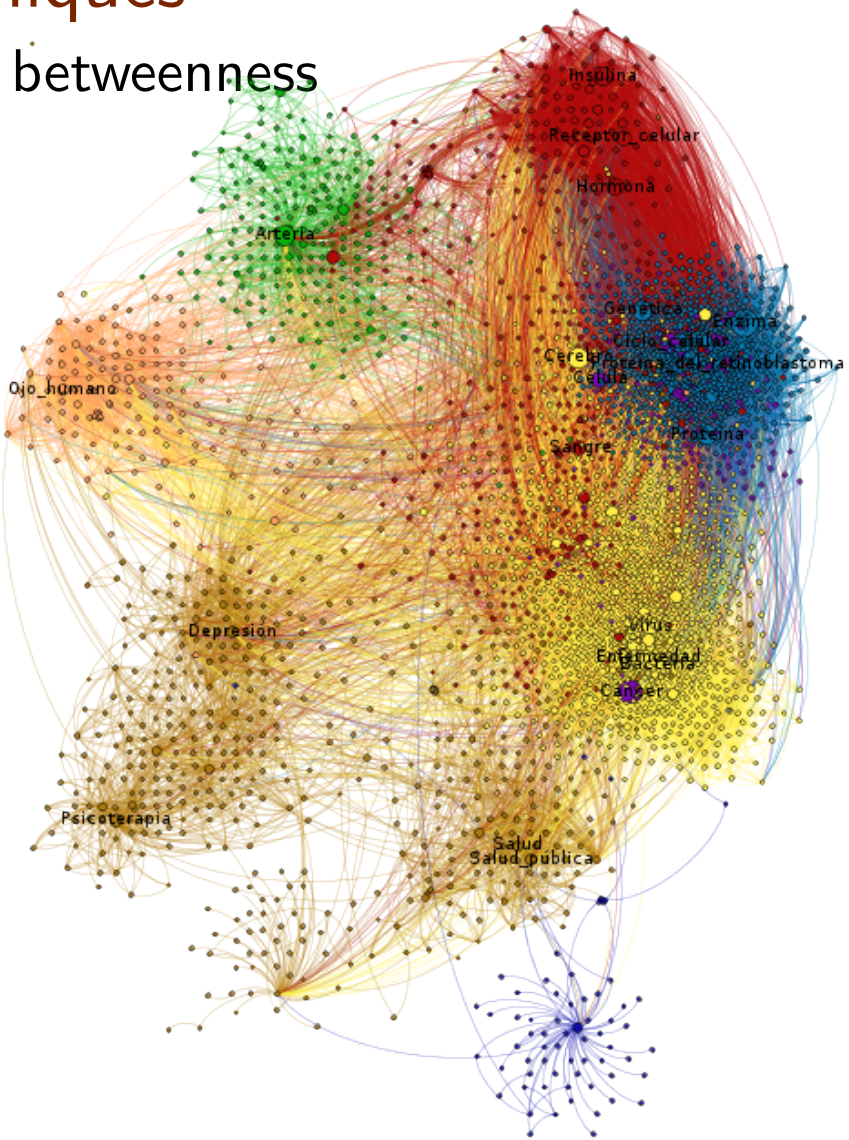
Table : Non Sci & Tech articles most linked from Life Sc.

Alimento	38	Spin_(House)	10	Muerte_de_Wolfgang_Amade ...	7
Lácteo	30	Heavy_(House)	10	Popper	7
Ataque_de_pánico	22	Cacería_(House)	10	Humpty_Dumpty_(House)	7
Carne	20	Cristina_Fernández_de_K ...	10	Efecto_del_falso_consenso	7
Vampiro	17	Café	10	Siete_sermones_a_los_m...	7
Queso	17	Vicio	9	Teoría_del_arte	7
Raza_(clasificación_de ...	17	Jack_el_Destripador	9	La_Gioconda	7
Creencias_sobre_vampiros	17	Clueless_(House)	9	Dune:_la_batalla_de ...	7
Vino	16	Martes_con_mi_viejo ...	9	Kurt_Cobain	7
Teología_moral_católica	16	George_Washington	9	Libre_albedrío	7
Tópico_literario	16	Hamburguesa	9	Marciano_(La_guerra_de ...	7
Ciberacoso	15	Coca-Cola	9	Choque	7
Homosexualidad	15	Evidencia_(filosofía)	9	Cy-Gor_(cómic)	7
Distracciones_(House)	13	Historia_del_arte	9	Bebida_alcohólica	7
Doctrina_de_la_Iglesia_C...	13	Edgar_Allan_Poe	9	Pobreza	6
All_in_(House)	13	House_M._D.	9	Frank_Zappa	6
Ludwig_van_Beethoven	13	Bebida_energizante	8	Franz_Liszt	6
Experimentación_nazi_en ...	11	Homosexualidad_en ...	8	Persona_sin_hogar	6
Hombres_que_tienen_sexo ...	11	Medicina_deportiva_(House)	8	Catolicismo	6
Edad_de_Oro_del_islam	11	Bruce_Lee	8	Historia_del_chocolate	6
La_Marató_de_TV3	10	Científico_loco	8	Islam	6
La_era_de_las_máquinas ...	10	Epicuro	8	Purgatorio_(Divina_C...	6
Richard_Wagner	10	Flor_del_Yo	8	A_salvo	6
Lesbianismo	10	Romeo_y_Julieta	8	Riding_in_Cars_with_Boys	6
Occam's_razor	10	Hambre	7	Angelina_Jolie	6

Table : Non Sci & Tech arts. most linking to Life Sci. articles

Applying SNA Techniques

Life Sci. articles with betweenness



Future Work

how to move forward?

Future Work

Things we can do:

- identify articles about Sci & Tech
- identify clusters of scientific topics
- analyze relationships between disciplines
- also between scientific disciplines and the not Sci & Tech world

Future Work

Things we can do:

- of course, analyze the size of each discipline
- editing activity
- relevance of articles

Future Work

How to apply the theoretical model?

Indicators of scientific and technological culture: Wikipedia

Carlos G. Figuerola, Angel Zazo Rodríguez, José L. Alonso Berrocal

Universidad de Salamanca. Instituto ECyT
(*figue—zazo—berrocal@usal.es)*
http://reina.usal.es