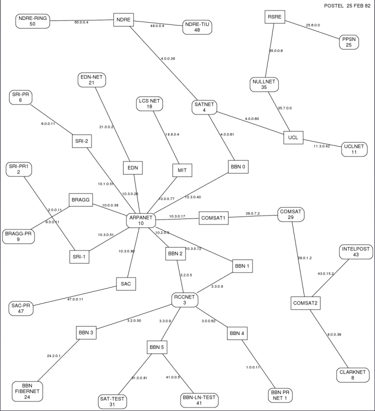
Analisi delle Reti Sociali
Network Evolution

Fosca Giannotti & Michele Berlingerio, KDDLab ISTI-CNR
http://kdd.isti.cnr.it/ *fosca.giannotti@isti.cnr.it, michele.berlingerio@isti.cnr.it*

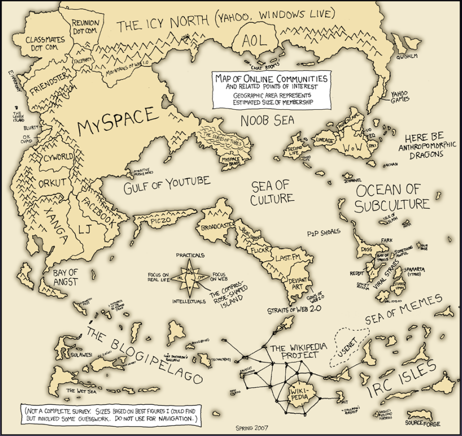
http://didawiki.cli.di.unipi.it/doku.php/dm/sna.ingegneria2011

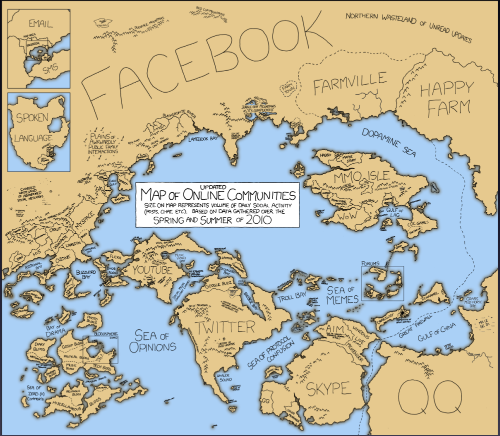
Networks evolve



internet in 1982.. ..and now!

Networks evolve





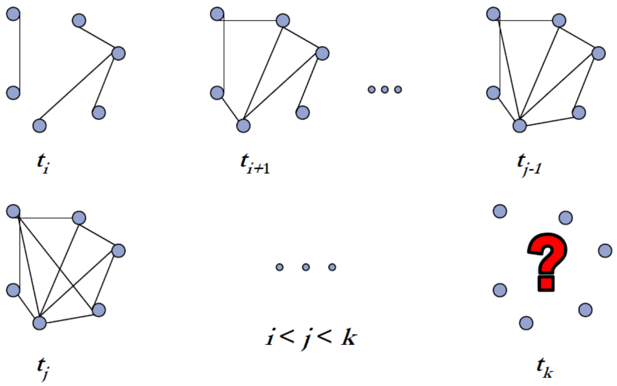
online communities in 2007 ..and in 2010

source: xkcd.com

Questions

- How does a network evolve over time?
- Is the evolution somehow regular?
- Can we predict new links?
- Is the evolution characterized by important *eras*?
- How do we find and characterize them?

Link Prediction



Given a snapshot of a social network at time t (or network evolution between t_1 and t_2), we seek to accurately predict the edges that will be added to the network during the interval from time t (or t_2) to a given future time t'.

LP - Applications

Overcoming the data-sparsity problem in recommender systems using collaborative filtering (Huang et al, 2005).

LP - Applications

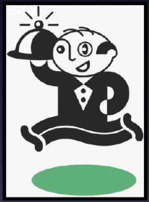
Identifying the structure of a criminal network
Predicting missing links in a criminal network using incomplete data.

LP - Applications

Accelerating a mutually beneficial professional- or academic connection that would have taken longer to form serendipitously (Farrell et al, 2005).

LP - Applications

To analyze users' navigation history to generate tools that increase navigational efficiency (Zhu 2003)
i.e. Predictive server prefetching



LP - Applications

Monitoring and controlling computer viruses that use email as a vector (Lim et al, 2005).

LP - Methods

- Assign a connection weight score(x, y) to pairs of nodes x, y, based on the input graph, and then produce a ranked list in decreasing order of score(x, y)
- Can be viewed as computing a measure of proximity or “similarity” between nodes x and y
- Supervised vs unsupervised

LP - Commong Neighbors

Newman 2001: The probability of scientists collaborating increases with the number of other collaborators they have in common.

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

LP - Jaccard Similarity

May be they have common neighbors because each one has a lot of neighbors, not because they are strongly related to each others

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

LP - Preferential Attachment

Newman 2001: The probability of co-authorship of x and y is correlated with the product of the number of collaborators of x and y

$$score(x, y) = |\Gamma(x)|.|\Gamma(y)|$$

LP - Adamic Adar

This gives more weight to neighbours that are not shared with many others.

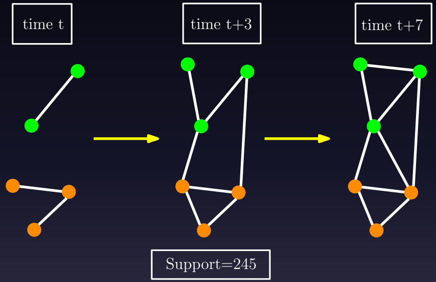
$$score(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(y)|}$$

LP - Comparisons

predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-two pairs)		*9.6*	*25.3*	*21.4*	*12.2*	*29.2*
common neighbors		**18.0**	**41.1**	**27.2**	**27.0**	**47.2**
preferential attachment		4.7	6.1	7.6	*15.2*	7.5
Adamic/Adar		*16.8*	**54.8**	**30.1**	**33.3**	**50.5**
Jaccard		*16.4*	**42.3**	19.9	**27.7**	*41.7*
SimRank	$\gamma = 0.8$	*14.6*	*39.3*	*22.8*	*26.1*	*41.7*
hitting time		6.5	23.8	*25.0*	3.8	13.4
hitting time—normed by stationary distribution		5.3	23.8	11.0	11.3	21.3
commute time		5.2	15.5	**33.1**	*17.1*	23.4
commute time—normed by stationary distribution		5.3	16.1	11.0	11.3	16.3
rooted PageRank	$\alpha = 0.01$	*10.8*	*28.0*	**33.1**	*18.7*	*29.2*
	$\alpha = 0.05$	*13.8*	*39.9*	**35.3**	*24.6*	*41.3*
	$\alpha = 0.15$	*16.6*	**41.1**	**27.2**	**27.6**	*42.6*
	$\alpha = 0.30$	*17.1*	**42.3**	*25.0*	**29.9**	*46.8*
	$\alpha = 0.50$	*16.8*	**41.1**	*24.3*	**30.7**	*46.8*
Katz (weighted)	$\beta = 0.05$	3.0	21.4	19.9	2.4	12.9
	$\beta = 0.005$	*13.4*	**54.8**	**30.1**	*24.0*	**52.2**
	$\beta = 0.0005$	*14.5*	**54.2**	**30.1**	**32.6**	**51.8**
Katz (unweighted)	$\beta = 0.05$	*10.9*	**41.7**	**37.5**	*18.7*	**48.0**
	$\beta = 0.005$	*16.8*	**41.7**	**37.5**	*24.2*	**49.7**
	$\beta = 0.0005$	*16.8*	**41.7**	**37.5**	*24.9*	**49.7**

Learning and Predicting the Evolution of a Network

Given n snapshots of an evolving network $G_1 \ldots G_n$ we want to mine patterns such as



to learn and predict the evolution of a network at the **local** level

Bringmann, Berlingerio, Bonchi, Gionis, IEEE Intelligent Systems 2010

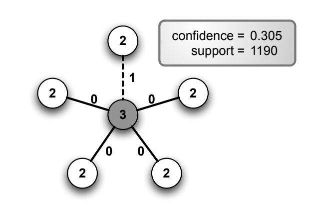
Learning and Predicting the Evolution of a Network

GERM, a new constraint-based frequent subgraph mining algorithm

Algorithm 1 *SubgraphMining*(G, S, s)

if $s \neq min(s)$ **then** return *// using our canonical form*
$S \leftarrow S \cup s$
enumerate all s' potential children with one edge growth
for all enumerated s' **do**
  // considering Δ *offset and our support definition*
  if $\sigma(s', G) \geq minSupp$ **then**
    SubgraphMining(G, S, s')
  end if
end for

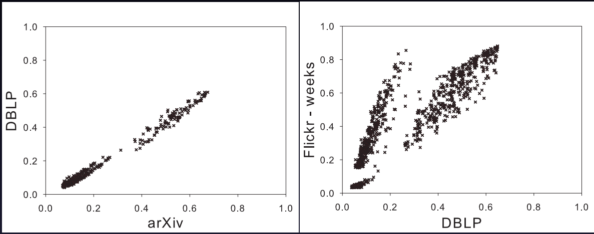
and get:



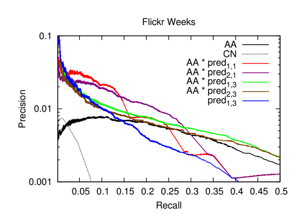
Learning and Predicting the Evolution of a Network

Results:

Rules characterize networks:



GERM-based prediction helps:



Discovery of Eras in Evolving Networks

Given n snapshots of an evolving network $G_1 \ldots G_n$ we want detected *eras* of evolution

- Cluster the snapshots at the global level
- Allow for evolution within one era
- Two eras characterized by different *speed* of evolution

Framework for Era Discovery

- Extraction of a time evolving network from real data
- Definition of a measure of dissimilarity among temporal snapshots of the same data
- Definition of clusters giving thresholds of such dissimilarity
- Merge of two (consecutive) clusters
- Assigning labels to clusters
- Realization of a dendrogram summarizing the clusters

Berlingerio, Coscia, Giannotti, Monreale, Pedreschi. M3SN2010 & PAKDD2010 & IDA Journal 2011

Dissimilarity measure

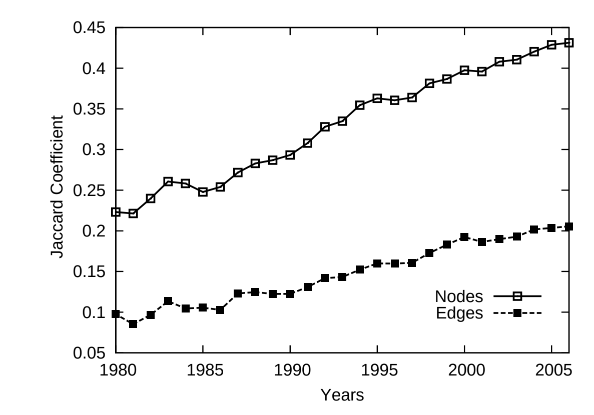
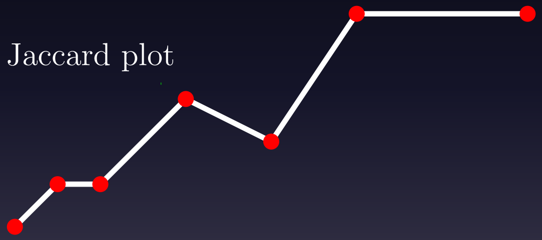


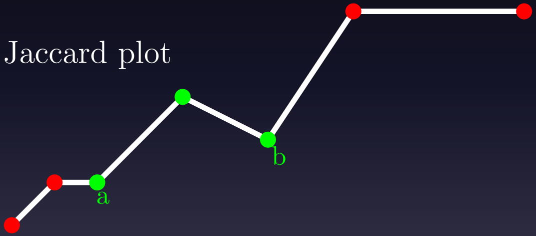
Figure: Evolution of the Jaccard Coefficient in DBLP

Dissimilarity measure

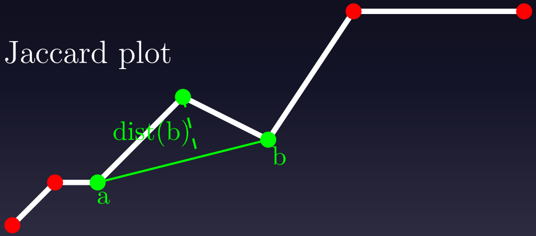


Dissimilarity measure

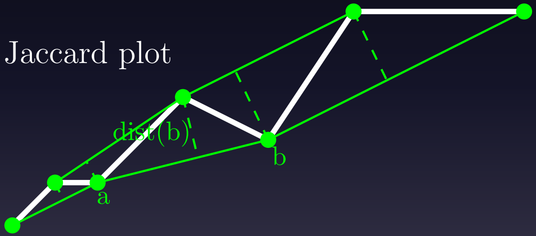
Jaccard plot



Dissimilarity measure



Dissimilarity measure



$$d(t_i, t_j) = \begin{cases} dist(t_{max(i,j)}) & \text{if } |i - j| = 1 \\ undefined & \text{otherwise} \end{cases}$$

Dissimilarity measure

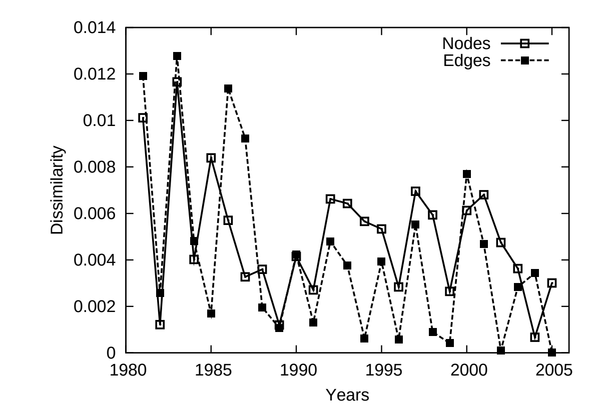
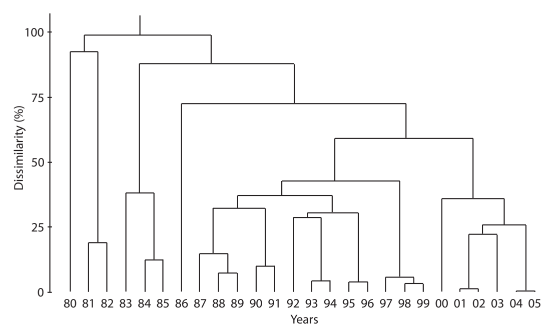


Figure: Dissimilarity Measure in DBLP

Eras on DBLP

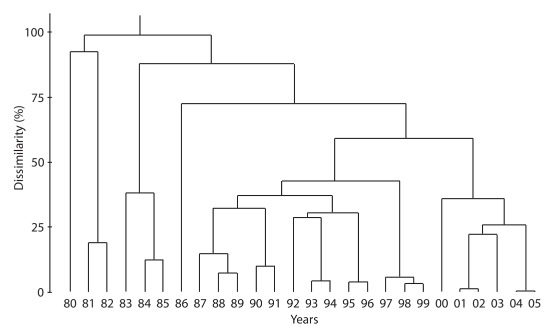
DBLP - edges



How to add semantic?

Eras on DBLP

DBLP - edges

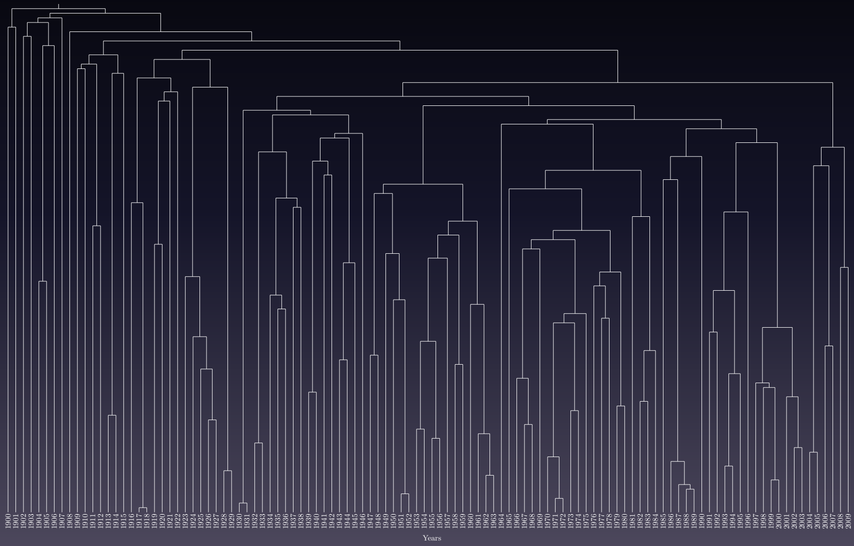


How to add semantic?

Labels assigned via TF/IDF

Start	End	Labels
1980	1982	pascal, language, database, data, micro-computer
1983	1985	prolog, database, online, abstract, expert
1987	1991	parallel, program, logic, abstract, database
1992	1996	parallel, program, logic, object oriented, computer
1997	1999	model, parallel, design, distributed, image
2001	2003	model, data, network, design, image
2004	2005	network, model, algorithm, web, data

Eras on IMDb


1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Years

Lessons learned..

- Network evolution is characterized by some regularity (evolution model)
- The network evolution model may be a sum of weaker signals
- The evolution model(s) may vary its/their speed (parameters)

Thank you!

Questions?