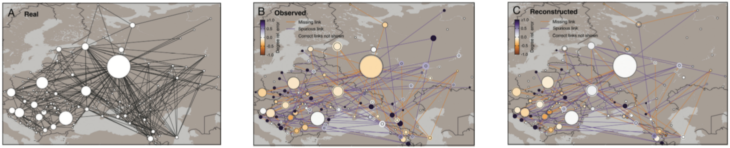
network *inference*

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2020/21

inference *overview*

- *inferring missing/spurious/hidden nodes/links*
  - due to *sampling*, *errors*, *noise* or *other* [GSP09, MBN15]
  - from *network structure*, *dynamics* or *other* [GRLK12]
- popular *predicting future links* that are likely to occur
  - recommendation of *friendship ties* on *Facebook* [BL11]
  - prediction of *product ratings* on *Amazon* [GLGMSP16]
  - prediction for costly *protein interaction* networks etc.



real, observed & reconstructed air transportation network [GSP09]

link *prediction*

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2020/21

prediction *overview*

which *links* are most *likely to occur*?

- link prediction by *local structure/dynamics*
  - *structural equivalence* [LW71] and *topological overlap* [RSM+02]
  - *node similarity* [LHN06] and *local dynamics* indices [ZLZ09]
- link prediction by *global structure/dynamics*
  - *regular equivalence* [WR83] and *link analysis* algorithms [JW02]
  - *community detection* [GN02] and *blockmodeling* [DBF05, Pei15]
- link prediction by *maximum likelihood* methods
  - *hierarchical* [CMN08] and *stochastic block* models [GSP09]
- link prediction by *probabilistic inference* methods
  - *probabilistic relational* models [FGKP99, SPH06]

prediction *equivalence*

links predicted by *structural equivalence*

— *common neighbors index* [LW71] for i and j is

$$s_{ij} = \sum_x A_{ix}A_{xj} = |\Gamma_i \cap \Gamma_j|$$

— *Jaccard neighbors index* [Jac01] for i and j is

$$s_{ij} = \frac{\sum_x A_{ix}A_{xj}}{\sum_x A_{ix} + \sum_x A_{xj} - \sum_x A_{ix}A_{xj}} = \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|}$$

— *Salton cosine similarity* [SM83] for i and j is

$$s_{ij} = \cos\theta_{ij} = \frac{\sum_x A_{ix}A_{xj}}{\sqrt{\sum_x A_{ix}^2}\sqrt{\sum_x A_{jx}^2}} = \frac{|\Gamma_i \cap \Gamma_j|}{\sqrt{k_i k_j}}$$

— *Leicht similarity index* [LHN06] for i and j is

$$s_{ij} = \frac{n\sum_x A_{ix}A_{xj}}{\sum_x A_{ix}\sum_x A_{jx}} = \frac{|\Gamma_i \cap \Gamma_j|}{k_i k_j / n} \approx \frac{|\Gamma_i \cap \Gamma_j|}{k_i k_j}$$

prediction *overlap*

links predicted by *topological overlap*

— *Sørensen neighbors index* [Sør48] for i and j is

$$s_{ij} = \frac{\sum_x A_{ix}A_{xj}}{\frac{1}{2}\left(\sum_x A_{ix} + \sum_x A_{xj}\right)} = \frac{|\Gamma_i \cap \Gamma_j|}{\frac{1}{2}(k_i + k_j)} \approx \frac{|\Gamma_i \cap \Gamma_j|}{k_i + k_j}$$

— *hub promoted index* [RSM+02] for i and j is

$$s_{ij} = \frac{\sum_x A_{ix}A_{xj}}{\min\left(\sum_x A_{ix}, \sum_x A_{xj}\right)} = \frac{|\Gamma_i \cap \Gamma_j|}{\min(k_i, k_j)}$$

— *hub depressed index* [LZ10] for i and j is

$$s_{ij} = \frac{\sum_x A_{ix}A_{xj}}{\max\left(\sum_x A_{ix}, \sum_x A_{xj}\right)} = \frac{|\Gamma_i \cap \Gamma_j|}{\max(k_i, k_j)}$$

prediction *models*

links predicted by *graph/network models*

— *configuration model index* [LHN06] for i and j is

$$s_{ij} = \frac{n\sum_x A_{ix}A_{xj}}{\sum_x A_{ix}\sum_x A_{jx}} = \frac{|\Gamma_i \cap \Gamma_j|}{k_i k_j / n} \approx \frac{|\Gamma_i \cap \Gamma_j|}{k_i k_j}$$

— *preferential attachment index* [BA99] for i and j is

$$s_{ij} = \sum_x A_{ix}\sum_x A_{xj} = k_i k_j$$

— *random graph index* [ER59] for i and j is

$$s_{ij} = \frac{\langle k \rangle}{n-1} \approx const.$$

prediction *dynamics*

links predicted by *local* *dynamics*

- *resource allocation index* [ZLZ09] for i and j is

$$s_{ij} = \sum_x \frac{A_{ix} A_{xj}}{\sum_y A_{xy}} = \sum_{x \in \Gamma_i \cap \Gamma_j} \frac{1}{k_x}$$

- *Adamic-Adar similarity index* [AA03] for i and j is

$$s_{ij} = \sum_x \frac{A_{ix} A_{xj}}{\log \sum_y A_{xy}} = \sum_{x \in \Gamma_i \cap \Gamma_j} \frac{1}{\log k_x}$$

- *WalkRank similarity index* [TFP06] for i and j is

$$p_i^t = \alpha \sum_{j \in \Gamma_i} p_j^t / k_j + (1 - \alpha) \delta_{it} \qquad s_{ij} = p_i^j + p_j^i$$

prediction *clusters*

links predicted by *node clusters*

— *community structure index* [YG11] for i and j is

- $\{C\}$ *communities* by *Louvain* [BGLL08] or *Infomap* [RB08]
- n_i and m_{c_i} *number of nodes* and *links* within C_i

$$s_{ij} = \begin{cases} \frac{m_{c_i}}{\binom{n_i}{2}} & \text{if } c_i = c_j \\ -\infty & \text{otherwise} \end{cases}$$

— *block model index* [HLL83, DBF05] for i and j is

- $\{C\}$ *clusters* by *stochastic block models* [Pei15]
- $m_{c_i c_j}$ *number of links* between C_i and C_j

$$s_{ij} = \begin{cases} \frac{m_{c_i}}{\binom{n_i}{2}} & \text{if } c_i = c_j \\ \frac{m_{c_i c_j}}{n_i n_j} & \text{otherwise} \end{cases}$$

prediction *framework*

link *prediction* as *ranking problem*

— *standard link prediction* setting

1. $L_N \leftarrow$ randomly *sample* $m/10$ *unlinked nodes* $\{i,j\} \notin L$
2. $L_P \leftarrow$ *remove* random $m/10$ *node links* $\{i,j\} \in L$
3. *compute* s_{ij} for $\{i,j\} \in L_N \cup L_P$ on *resulting* L

— *temporal link prediction* setting

1. $L_N \leftarrow$ randomly *sample* $|L_P|$ *unlinked nodes* $\{i,j\} \notin L$
2. $L_P \leftarrow$ *remove node links* $\{i,j\} \in L$ *after time* t
3. *compute* s_{ij} for $\{i,j\} \in L_N \cup L_P$ on L *at time* t

— *Pearson/Spearman correlation* or *AUC measure*

$$[\overbrace{0\ 0\ 0\ \ldots\ 0\ 0\ 0}^{\text{ideal } s_{ij} \text{ for } L_N}\ \overbrace{1\ 1\ 1\ \ldots\ 1\ 1\ 1}^{\text{ideal } s_{ij} \text{ for } L_P}]$$

prediction *scale-free*

— *link prediction* in synthetic *scale-free graph* [BA99]

1st *highest AUC* by *stochastic block models* [Pei15]

2nd *highest AUC* by *preferential attachment* [BA99]

class	index	Pearson	Spearman	AUC
models	*preferential*	0.128	0.347	0.701
equivalence	neighbors	0.105	0.135	0.530
	Jaccard	0.019	0.131	0.529
	Salton	0.043	0.131	0.529
	Leicht	−0.008	0.131	0.529
dynamics	allocation	0.091	0.135	0.530
	Adamic-Adar	0.104	0.135	0.530
clusters	modularity	0.002	0.005	0.502
	map equation	0.009	0.034	0.503
	block model	0.168	0.370	0.711
baseline	random	−0.001	−0.001	0.499

prediction *small-world*

— *link prediction* in synthetic *small-world graph* [WS98]

1st *highest AUC* by *stochastic block models* [Pei15]

2nd *highest AUC* by *common neighbors index* [LW71]

class	**index**	**Pearson**	**Spearman**	**AUC**
models	preferential	−0.563	−0.547	0.187
equivalence	*neighbors*	0.721	0.786	0.903
	Jaccard	0.686	0.785	0.903
	Salton	0.729	0.785	0.902
	Leicht	0.730	0.785	0.903
dynamics	*allocation*	0.719	0.785	0.902
	Adamic-Adar	0.720	0.785	0.902
clusters	modularity	0.743	0.754	0.885
	map equation	0.643	0.649	0.807
	block model	0.737	0.754	0.931
baseline	random	−0.003	−0.002	0.499

prediction *human*

— *link prediction* in *human protein interaction* map

1st *highest AUC* by *stochastic block models* [Pei15]

2nd *highest AUC* by *preferential attachment* [BA99]

class	**index**	**Pearson**	**Spearman**	**AUC**
models	*preferential*	0.231	0.719	0.915
equivalence	neighbors	0.342	0.676	0.845
	Jaccard	0.301	0.648	0.830
	Salton	0.391	0.646	0.830
	Leicht	−0.005	0.625	0.819
dynamics	allocation	0.291	0.681	0.847
	Adamic-Adar	0.343	0.680	0.847
clusters	modularity	0.284	0.381	0.672
	map equation	0.220	0.408	0.660
	block model	0.344	0.746	0.929
baseline	random	0.000	0.000	0.500

prediction *P2P*

— *link prediction* in *P2P file transfer* network [LKF07]

1st *highest AUC* by *stochastic block models* [Pei15]

2nd *highest AUC* by *preferential attachment* [BA99]

class	**index**	**Pearson**	**Spearman**	**AUC**
models	*preferential*	0.379	0.378	0.717
equivalence	neighbors	0.113	0.120	0.515
	Jaccard	0.093	0.120	0.515
	Salton	0.098	0.120	0.515
	Leicht	0.055	0.120	0.515
dynamics	allocation	0.087	0.120	0.515
	Adamic-Adar	0.102	0.120	0.515
clusters	modularity	0.081	0.121	0.531
	map equation	0.096	0.113	0.513
	block model	0.487	0.621	0.837
baseline	random	−0.002	−0.002	0.499

prediction *iMDB*

— *link prediction* in *iMDB collaboration* network [BA99]

1st *highest AUC* by *stochastic block models* [Pei15]

2nd *highest AUC* by *resource allocation index* [ZLZ09]

class	**index**	**Pearson**	**Spearman**	**AUC**
models	preferential	0.359	0.589	0.840
equivalence	*neighbors*	0.491	0.875	0.970
	Jaccard	0.609	0.876	0.970
	Salton	0.724	0.877	0.970
	Leicht	0.355	0.869	0.967
dynamics	*allocation*	0.627	0.878	0.971
	Adamic-Adar	0.520	0.876	0.970
clusters	modularity	0.345	0.826	0.948
	map equation	0.421	0.785	0.909
	block model	0.544	0.856	0.986
baseline	random	−0.003	−0.003	0.498

prediction *nd.edu*

— *link prediction* in *nd.edu* *web* graph [BA99]

1st *highest AUC* by *modularity optimization* [BGLL08]

2nd *highest AUC* by *resource allocation index* [ZLZ09]

class	index	Pearson	Spearman	AUC
models	preferential	0.094	0.548	0.816
equivalence	*neighbors*	0.346	0.717	0.855
	Jaccard	0.453	0.716	0.854
	Salton	0.526	0.716	0.854
	Leicht	0.257	0.715	0.854
dynamics	*allocation*	0.181	0.718	0.855
	Adamic-Adar	0.334	0.718	0.855
clusters	*modularity*	0.197	0.767	0.893
	map equation	0.391	0.703	0.844
	block model	-	-	-
baseline	random	−0.001	−0.001	0.499

prediction *WoS*

— *link prediction* in *WoS citation* network [ŠF17]

1st *highest AUC* by *modularity optimization* [BGLL08]

2nd *highest AUC* by *common neighbors index* [LW71]

class	index	Pearson	Spearman	AUC
models	preferential	0.082	0.509	0.794
equivalence	*neighbors*	0.434	0.754	0.880
	Jaccard	0.499	0.753	0.880
	Salton	0.574	0.753	0.880
	Leicht	0.258	0.753	0.880
dynamics	*allocation*	0.449	0.754	0.880
	Adamic-Adar	0.454	0.754	0.880
clusters	*modularity*	0.082	0.779	0.908
	map equation	0.392	0.546	0.734
	block model	-	-	-
baseline	random	0.000	0.000	0.500

prediction *Texas*

— *link prediction* in *Texas road* map [LLDM09]

1st *highest AUC* by *modularity optimization* [BGLL08]

2nd *highest AUC* by *map equation method* [RB08]

class	index	Pearson	Spearman	AUC
models	*preferential*	−0.353	−0.311	0.322
equivalence	neighbors	0.230	0.233	0.551
	Jaccard	0.217	0.232	0.551
	Salton	0.225	0.232	0.551
	Leicht	0.202	0.232	0.551
dynamics	allocation	0.225	0.232	0.551
	Adamic-Adar	0.225	0.232	0.551
clusters	*modularity*	0.060	0.736	0.868
	map equation	0.335	0.362	0.616
	block model	-	-	-
baseline	random	0.000	0.000	0.500

inference *references*

Lada A Adamic and Eytan Adar.
Friends and neighbors on the Web.
Soc. Networks, 25(3):211–230, 2003.

A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286(5439):509–512, 1999.

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.
Fast unfolding of communities in large networks.
J. Stat. Mech., P10008, 2008.

Tom Britton and David Lindenstrand.
Inhomogeneous epidemics on weighted networks.
e-print arXiv:11124718v1, 2011.

Aaron Clauset, Cristopher Moore, and M. E. J. Newman.
Hierarchical structure and the prediction of missing links in networks.
Nature, 453(7191):98–101, 2008.

Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj.
Generalized Blockmodeling.
Cambridge University Press, Cambridge, 2005.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj.
Exploratory Social Network Analysis with Pajek: Expanded and Revised Second Edition.
Cambridge University Press, Cambridge, 2011.

inference *references*

David Easley and Jon Kleinberg.
Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

Ernesto Estrada and Philip A. Knight.
A First Course in Network Theory.
Oxford University Press, 2015.

P. Erdős and A. Rényi.
On random graphs I.
Publ. Math. Debrecen, 6:290–297, 1959.

Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer.
Learning probabilistic relational models.
In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1300–1309, Stockholm, Sweden, 1999.

Antonia Godoy-Lorite, Roger Guimerà, Cristopher Moore, and Marta Sales-Pardo.
Accurate and scalable social recommendation using mixed-membership stochastic block models.
P. Natl. Acad. Sci. USA, 113(50):14207–14212, 2016.

M. Girvan and M. E. J Newman.
Community structure in social and biological networks.
P. Natl. Acad. Sci. USA, 99(12):7821–7826, 2002.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause.
Inferring networks of diffusion and influence.
ACM Trans. Knowl. Discov. Data, 5(4):21, 2012.

inference *references*

Roger Guimerà and Marta Sales-Pardo.
Missing and spurious interactions and the reconstruction of complex networks.
P. Natl. Acad. Sci. USA, 106(52):22073–22078, 2009.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt.
Stochastic blockmodels: First steps.
Soc. Networks, 5(2):109–137, 1983.

Paul Jaccard.
Étude comparative de la distribution florale dans une portion des Alpes et des Jura.
Bulletin del la Société Vaudoise des Sciences Naturelles, 37:547–579, 1901.

G. Jeh and J. Widom.
SimRank: A measure of structural-context similarity.
In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

E. A. Leicht, Petter Holme, and M. E. J. Newman.
Vertex similarity in networks.
Phys. Rev. E, 73(2):026120, 2006.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos.
Graph evolution: Densification and shrinking diameters.
ACM Trans. Knowl. Discov. Data, 1(1):1–41, 2007.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney.
Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters.
Internet Math., 6(1):29–123, 2009.

inference *references*

F. Lorrain and H. C. White.
Structural equivalence of individuals in social networks.
J. Math. Sociol., 1(1):49–80, 1971.

Linyuan Lü and Tao Zhou.
Link prediction in complex networks: A survey.
Physica A, 2010.

Travis Martin, Brian Ball, and M. E. J. Newman.
Structural inference for uncertain networks.
e-print arXiv:150605490v1, 2015.

Mark E. J. Newman.
Networks.
Oxford University Press, Oxford, 2nd edition edition, 2018.

Tiago P. Peixoto.
Model selection and hypothesis testing for large-scale network models with overlapping groups.
Phys. Rev. X, 5(1):011033, 2015.

M. Rosvall and C. T. Bergstrom.
Maps of random walks on complex networks reveal community structure.
P. Natl. Acad. Sci. USA, 105(4):1118–1123, 2008.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and Albert László Barabási.
Hierarchical organization of modularity in metabolic networks.
Science, 297(5586):1551–1555, 2002.

Lovro Šubelj and Dalibor Fiala.
Publication boost in Web of Science journals and its effect on citation distributions.
J. Assoc. Inf. Sci. Tec., 68(4):1018–1023, 2017.

G. Salton and M. J. McGill.
Introduction to Modern Information Retrieval.
McGraw-Hill, 1983.

Thorvald Julius Sørensen.
A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.
Biol. Skr., 5:1–34, 1948.

B. Schölkopf, J. Platt, and T. Hofmann.
Stochastic relational models for discriminative link prediction.
In *Proceedings of the Neural Information Processing Systems Conference*, pages 1553–1560, 2006.

H. Tong, Christos Faloutsos, and Jia-Yu Pan.
Fast random walk with restart and its applications.
In *Proceedings of the IEEE International Conference on Data Mining*, pages 613–622, Washington, DC, USA, 2006.

D. R. White and K. P. Reitz.
Graph and semigroup homomorphisms on networks of relations.
Soc. Networks, 5(2):193–234, 1983.

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.

Bowen Yan and Steve Gregory.
Finding missing edges and communities in incomplete networks.
J. Phys. A: Math. Theor., 44(49):495102, 2011.

inference *references*

Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang.
Predicting missing links via local information.
Eur. Phys. J. B, 71(4):623–630, 2009.