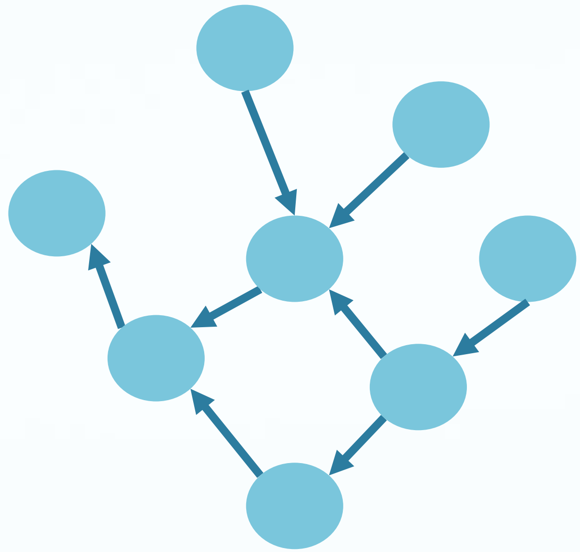
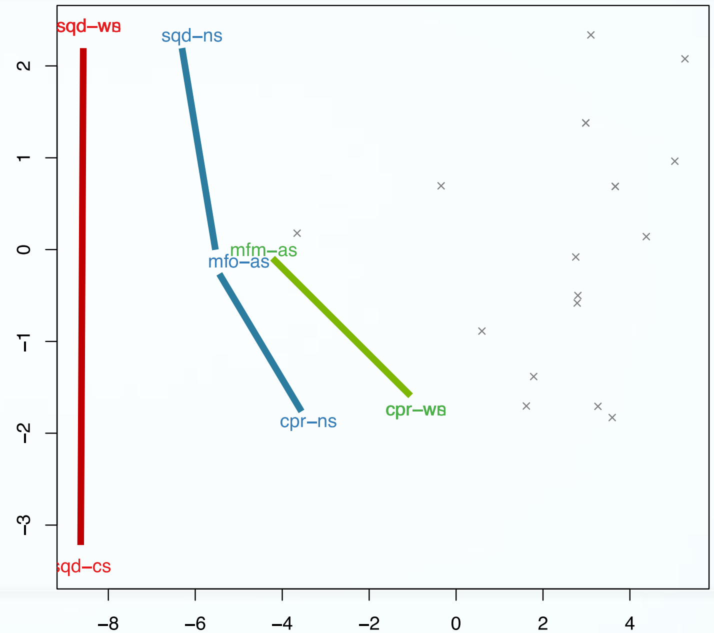




Relevance Network Approach to Network Reconstruction

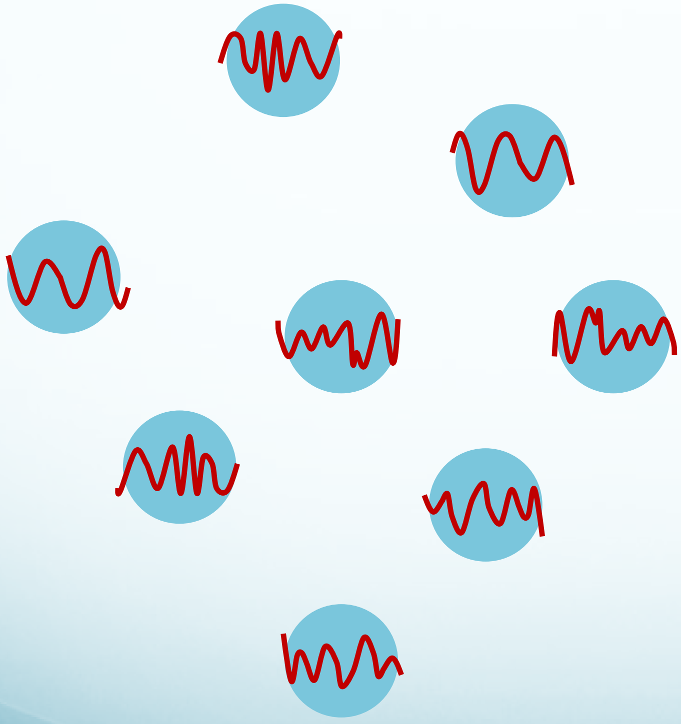
Ljupčo Todorovski
University of Ljubljana, Faculty of Administration, Slovenia

Vladimir Kuzmanovski, Sašo Džeroski
Jožef Stefan Institute, Dept of Knowledge Technologies, Slovenia

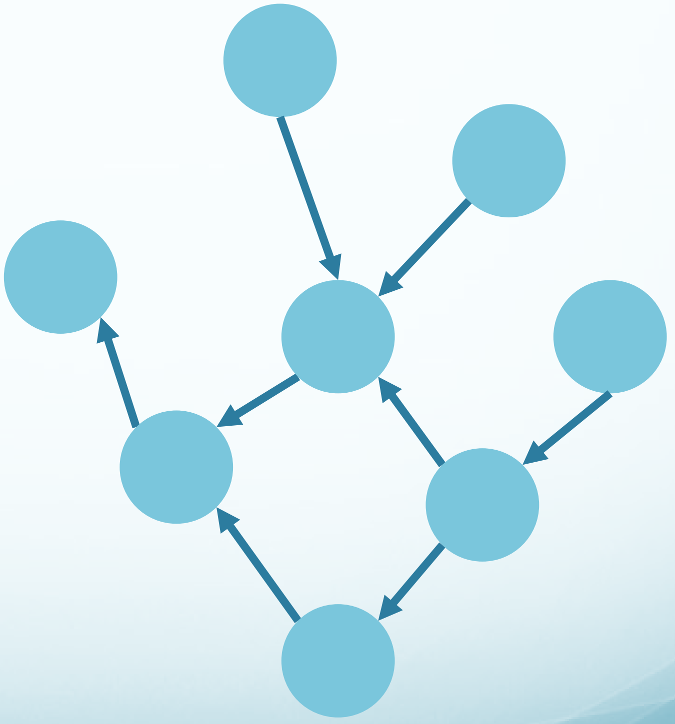
Network Reconstruction Task

- Also network inference task

Given time-series data



Find network links

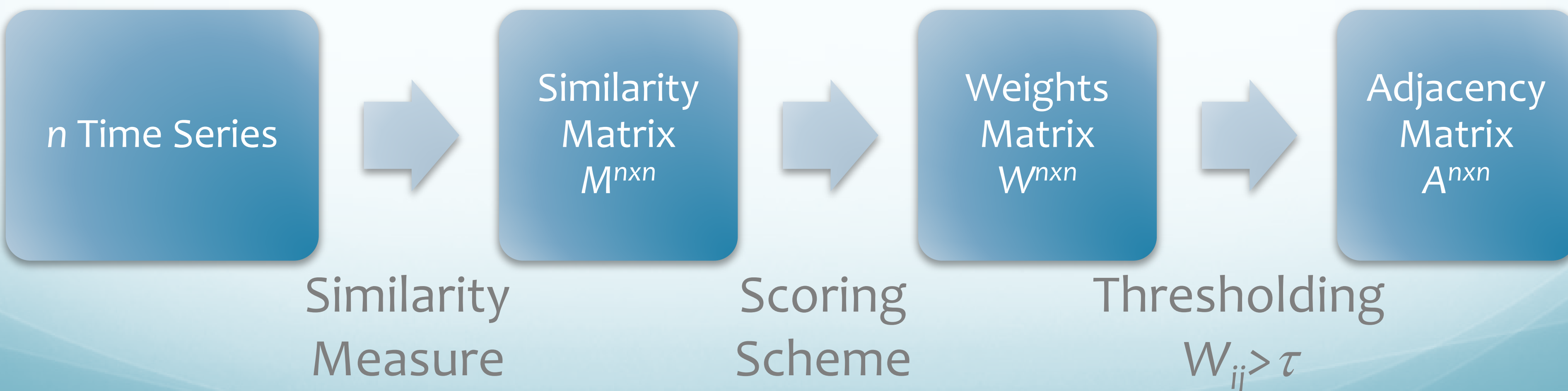


Applications and Methods

- Example applications by domain
  - Bioinormatics: from expression data to gene regulatory networks
  - Social networks: from time-series of number of retweets to Twitter influence networks
  - Collaborative environments: from number of article edits to information propagation networks in Wikipedia
  - Climate: from time-series data measured at a regular grid over the globe, identify geographical regions affected by El Nino
- Methods and approaches to network reconstruction
  - Operate on various target formal representation of the networks
  - Methods for Bayesian networks, more general graphical models
  - This talk: Relevance Network Approach

Relevance Network Approach

- Assumption: (high) similarity between the time series observed in two nodes indicate a presence of network link between them
  - Thus: the focus is on **measuring similarity** between time series
  - Problem: **Similarity** often **symmetric**, leading to undirected nets
  - Solution: **symmetry-breaking** scoring schemes



What is this Talk About?

- Brief **survey of the relevance network approaches**
  - Similarity measures and scoring schemes
  - Spoiler alert: in sum, **there are (too) many of them**
- So: which similarity measure and scoring scheme should be used?
  - Michelangelo's answer: **all of them at the same time**
  - We rephrase the question into: **What works where?**
- Ideally, we would be able to **provide recommendations**
  - You should use similarity measure X and scoring scheme Y, since
  - There is a large number of nodes in the network, and
  - The time series are long

Talk Outline

- Introduction and motivation
- Relevance network (RN) approach
  - Similarity measures
  - Scoring schemes
- Empirical comparison of the RN variants
  - Experimental setup: networks, data sets, performance measures
  - Comparison methodology
  - Empirical results: what works where?
- Conclusion and further work

Similarity Measures (SM)

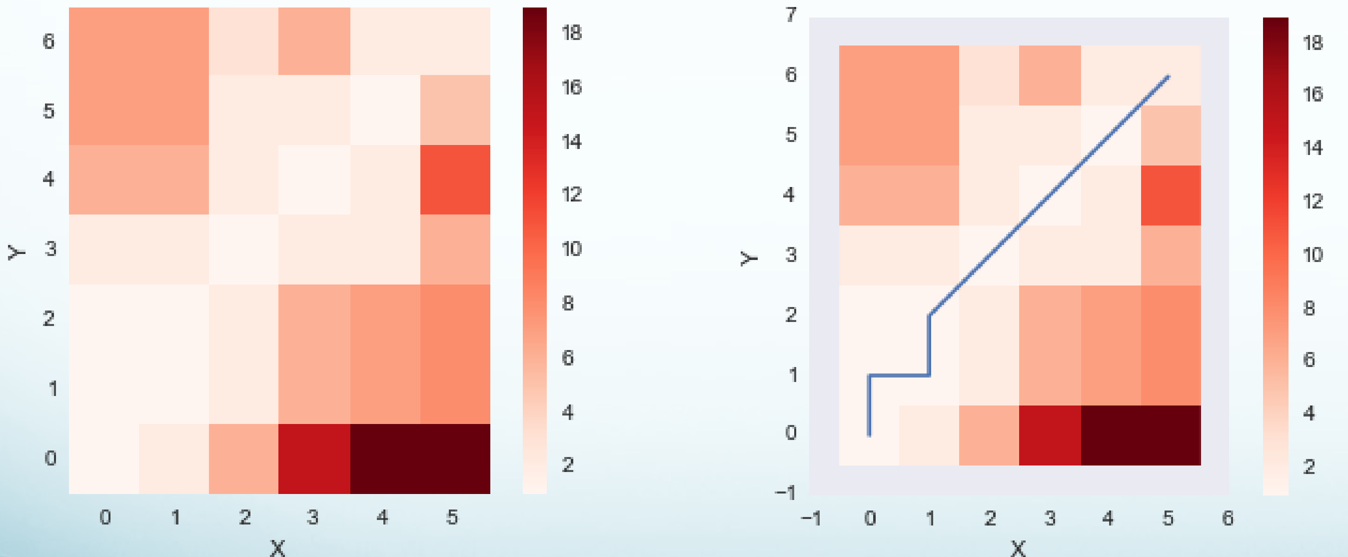
- Similarity measure $m: R^n \times R^n \rightarrow R$
  - Detects (non)linear relation between two given time series
- Many different measures proposed; can be clustered in 5 classes
  - Distances
  - Dynamic Time Warping
  - Correlations
  - Mutual Information
  - Symbolic

SM: Distances (Norms)

- Distance-based similarities **regard time series as vectors**
- Distance between x and y defined as a p-norm of the vector $x-y$:
  - $d_p(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/p}$
  - $p=1$: Manhattan distance
  - $p=2$: Euclidian distance
  - Often used (please do not ask why) $p=10$
- From **distance to similarity**?
  - Many ways, most simple $m(x,y) = -d_p(x,y)$
  - Or, if you are afraid of negative numbers $m(x,y) = 1 / d_p(x,y)$

SM: Dynamic Time Warping

- Optimal mapping between two time series *x* and *y*, such that
  - Points from *x* are linked to points in *y*
  - Each point should participate in at least one link
  - The sum of the link lengths is minimal



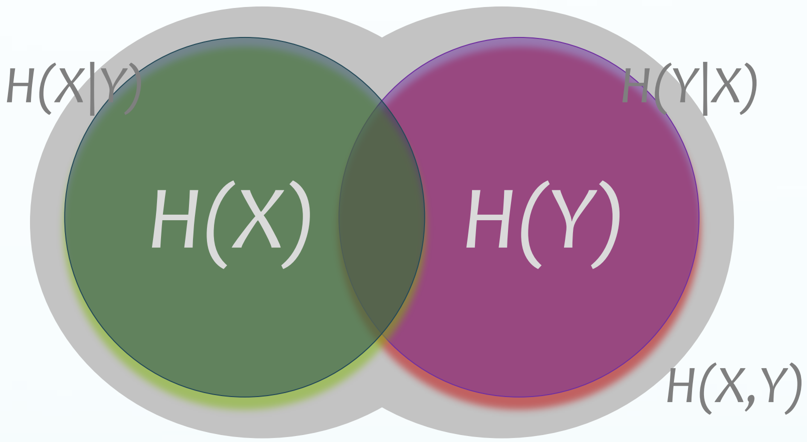
- Finding the optimal mapping: **dynamic programming** formula
  - **Different variants of the formula** lead to **different DTW measures**

SM: Correlation Coefficients

- Regard time series as random variables X and Y
  - Pearson $r_P(X,Y) = E[(X-E[X])(Y-E[Y])] / (E[(X-E[X])^2]\ E[(Y-E[Y])^2])$
- More robust to non-normal distributions
  - Spearman $r_S(X,Y) = r_P(ranks(X), ranks(Y))$
  - Kendall $r_K(X,Y) = 2(n_C - n_D) / (n\ (n-1))$
    - n_C: number of concordant pairs of time points
    - n_D: number of dis-concordant pairs of time points
- Often squared values used
  - Since we are not inferring the *direction of the relationship* (positive, negative), but only to its degree
  - We are **not referring here to the causal direction**, which could have been interpreted as a link direction

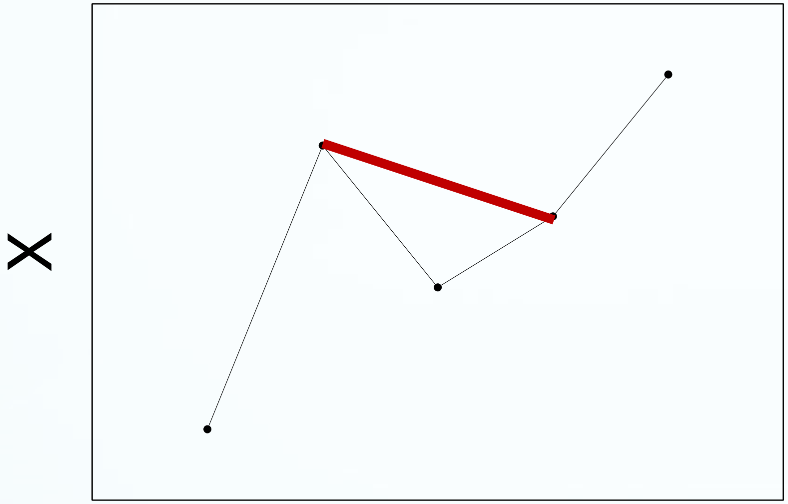
SM: Mutual Information

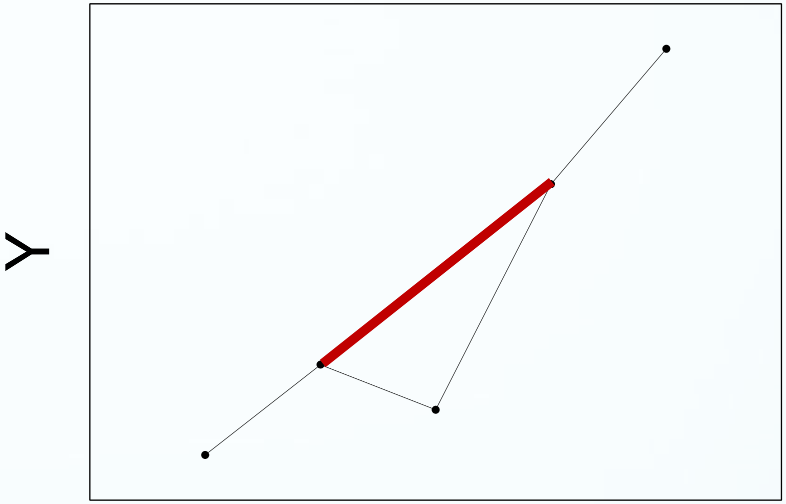
- Treat the time series as random variables X and Y
  - $MI(X,Y) = H(X) + H(Y) - H(X,Y)$, where H denotes entropy



- Requires discretization of the numeric variables; hence **different variants** corresponding to **different discretization methods**
  - Equal-frequency or equal-width bins
  - Various techniques for determining the number of bins

SM: Simple Qualitative/Symbolic Distance



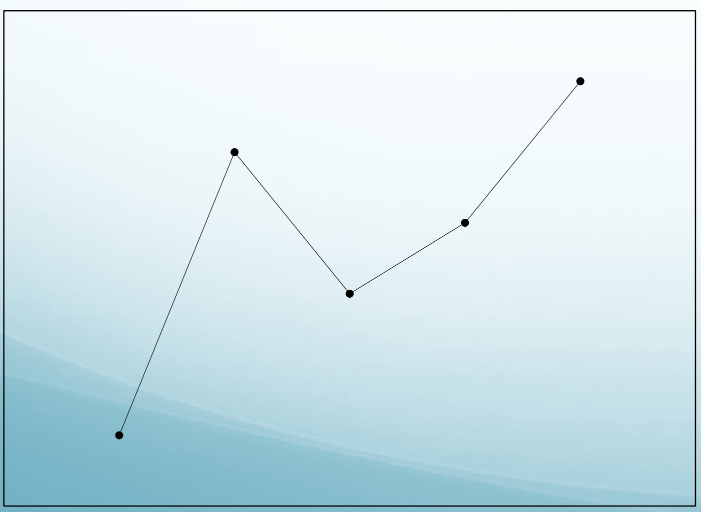


- Comparing simple pairwise increase/decrease trends
  - (t_1,t_2): X↑Y↑, (t_1,t_3): X↑Y↑, (t_1,t_4): X↑Y↑, (t_1,t_5): X↑Y↑
  - (t_2,t_3): X↓Y↓, (t_2,t_4): X↓Y↑, (t_2,t_5): X↑Y↑
  - (t_3,t_4): X↑Y↑, (t_3,t_5): X↑Y↑
  - (t_4,t_5): X↑Y↑
- 1 difference in 10 pairwise comparisons: $d(X,Y) = 1/10 = 0.1$

SM: Symbolic Dynamics

- Transformation of time series to a vector of order patterns
  - Calculating **distances** or **mutual information** on **symbolic vectors**

Order patterns for 3 time points					
					
P_1	P_2	P_3	P_4	P_5	P_6



$(t_1,t_2,t_3){:}P_2$ $(t_1,t_2,t_4){:}P_2$ $(t_1,t_2,t_5){:}P_1$ $(t_1,t_3,t_4){:}P_1$ $(t_1,t_3,t_5){:}P_1$ $(t_1,t_3,t_5){:}P_1$

$(t_2,t_3,t_4){:}P_4$ $(t_2,t_3,t_5){:}P_5$ $(t_2,t_4,t_5){:}P_5$

$(t_3,t_4,t_5){:}P_1$

Symbolic vector $(P_2,P_2,P_1,P_1,P_1,P_1,P_4,P_5,P_5,P_1)$

Symmetry Breaking Scoring Schemes

- Time shifting (TS)
  - Common way to **infer** the **directionality of causal relationships**
  - Observing the trend of **correlation change when shifting one time series,** provides a hint on the direction of the causal relationship
  - ***X*→*Y*: shifting *X* (the cause) to the right** (forward in time) will **increase the similarity/correlation** between *X* and *Y*
- Asymmetric Weighting (AWE)
  - Similarity matrix elements divided by the sum of the elements in the corresponding column, i.e., $W_{ij} = M_{ij} / \Sigma_k M_{kj}$
  - Can be used alone or in combination with TS

Other Scoring Schemes

- **Must be combined with time shifting** to identify causal direction
- Context Likelihood of Relatedness (CLR)
  - Uses the distribution of the values in the matrix *M* for
  - Normalization using the averages and standard deviations of the values in the columns and and rows of *M*
- Identifying and discriminating indirect links
  - Algs for Reconstruction of Accurate Cellular Networks (ARACNE)
  - Heuristic for identification of indirect links: $M_{ik} <= min(M_{ij}, M_{jk})$
  - Maximum Relevance / minimum redundancy Network (MRNET)
  - Assigns higher ranks to direct links, lower ranks to indirect links

Talk Outline

- Introduction and motivation
- Relevance network (RN) approach
  - Similarity measures
  - Scoring schemes
- Empirical comparison of the RN variants
  - Experimental setup: networks, data sets, performance measures
  - Comparison methodology
  - Empirical results: what works where?
- Conclusion and further work

Networks and Data: Yeast

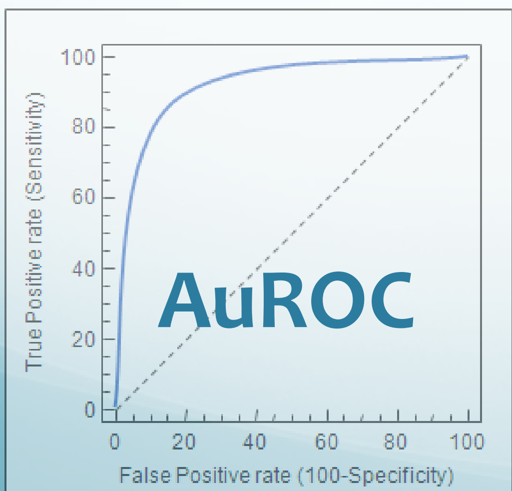
- Four Yeast networks (size: #nodes, #links, density)
  - YN1: 42, 61, 1.1E-2
  - YN2: 75, 135, 4.1E-3
  - YN3: 300, 448, 1.5E-3
  - YN4: 188, 283, 2.4E-3
- 13 time-series data sets that only partially cover network nodes
  - Real measurements that only **partially cover network nodes**
  - For each network **data sets** selected that **cover at least 95% nodes**
  - YN1: 6 data sets, YN2: 2, YN3: 5, and YN4: 3 data sets
- Total of 20 network reconstruction tasks

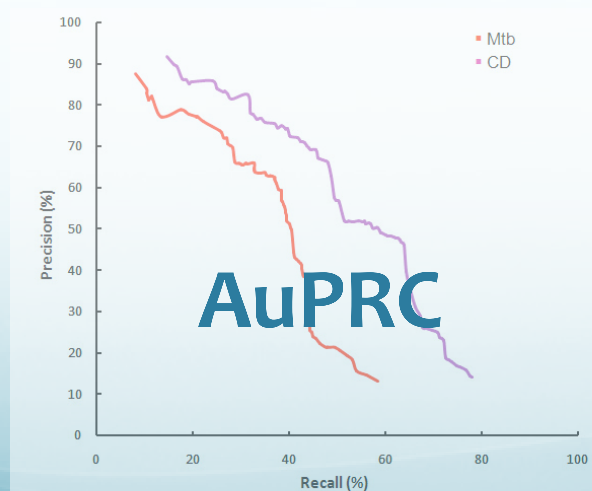
Networks and Data: Dream5

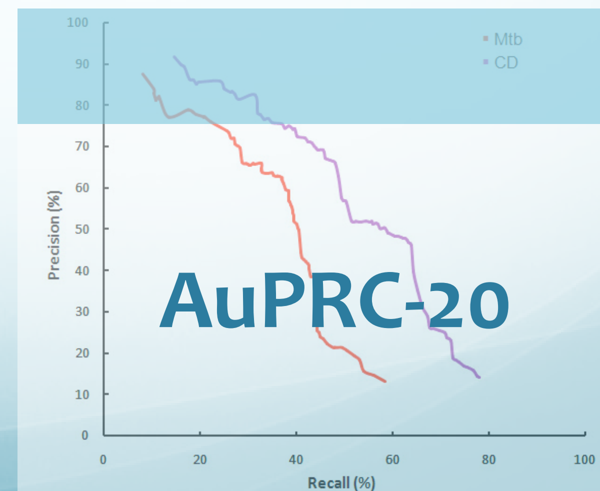
- Two Dream5 NR-challenge networks (size: #nodes, #links, density)
  - DN1: 4511, 2066, 1.1E-03
  - DN2: 5950, 3940, 3.8E-04
- Four synthetic (simulated) data sets that cover all network nodes
  - DN1: 2 data sets, DN2: 2 data sets
- Total of 4 network reconstruction tasks

Methods and Performance Measures

- 114 (=19*6) Relevance Network Approach Variants
  - 19 similarity measures: 3 distances, 3 DTW variants, 3 correlation coefficients, 4 mutual-information variants, 6 symbolic variants
  - 6 scoring schemes: TS, AWE, AWE+TS, CLR+TS, ARACNE+TS, MRNET+TS
- 2,736 (=114*(20+4)) experiments
- Three performance measures



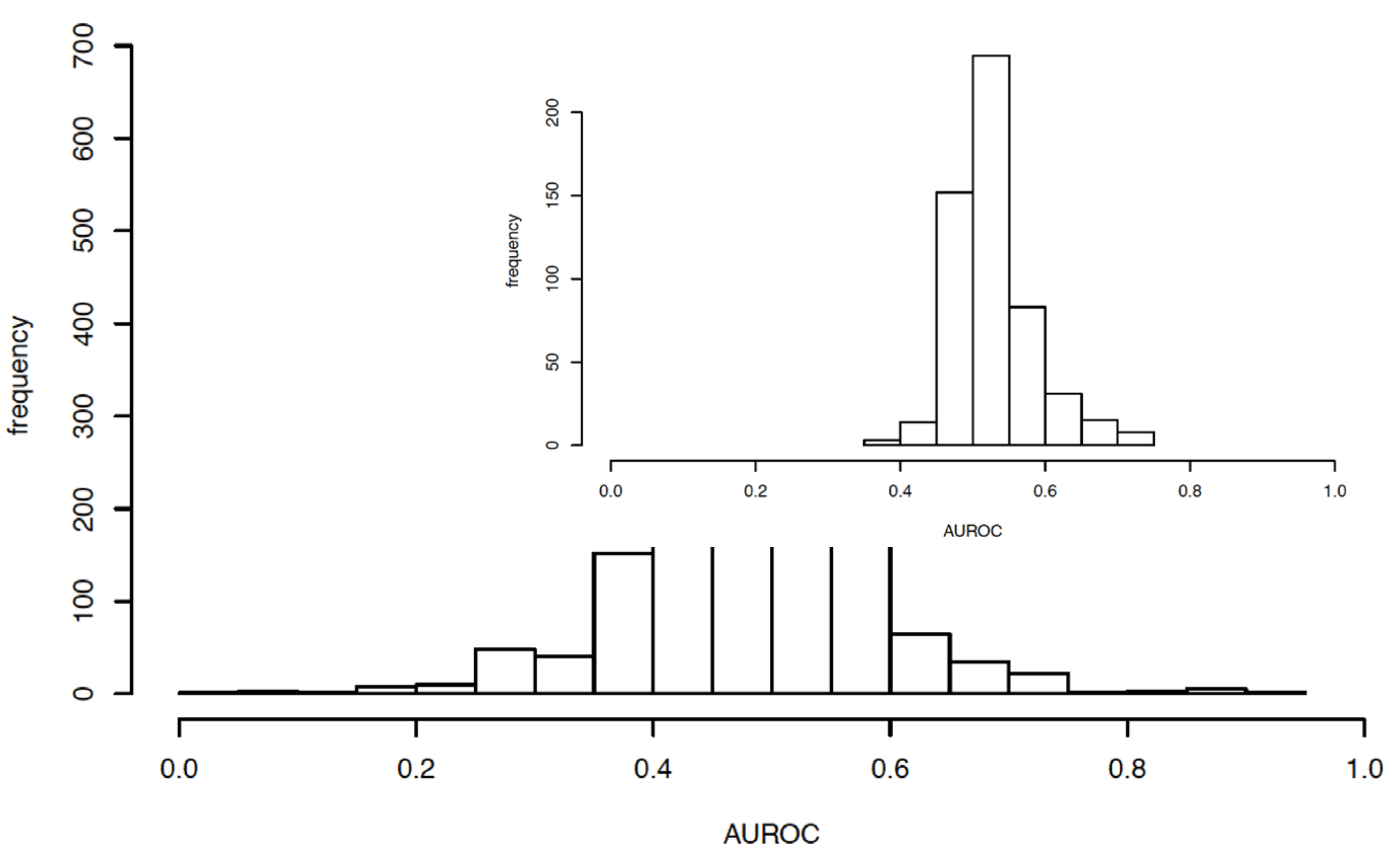




Comparison Methodology

- One sample Student's t-test (p-value < 0.05)
  - Identify well-performing methods that **on average perform significantly better than the default**/random NR
  - **WRT at least one performance measure:** AUROC default 0.5, AUPRC default 0.5, AUPRC-20% default 0.1
  - The average calculated on the 20 network reconstruction problems
- Compare the average rankings of the well-performing methods
  - Pareto fronts in the 3D performance space
  - Observing method ranks (can be also performances)
- Which methods are in the first three Pareto fronts:
  - Similarity measures? Scoring schemes?

Methods Selection: T-Test



Pareto Fronts in the Performance Space

6.3–13.3

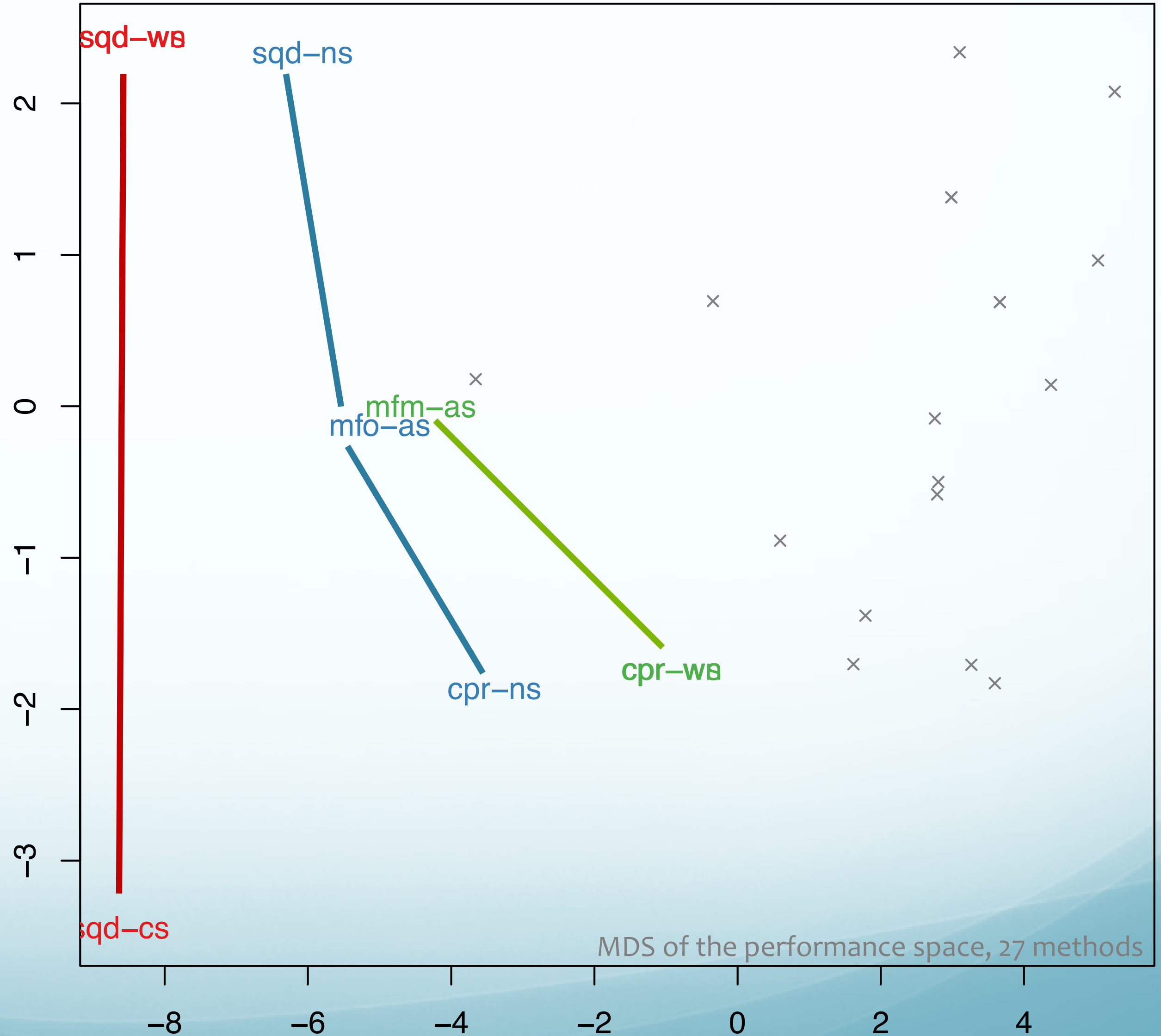
SymQD-AWE
SymQD-AWE-TS
SymQD-CLR-TS

8.1–13.8

CorrP-TS
MI-ARACNE-TS
SymQD-TS

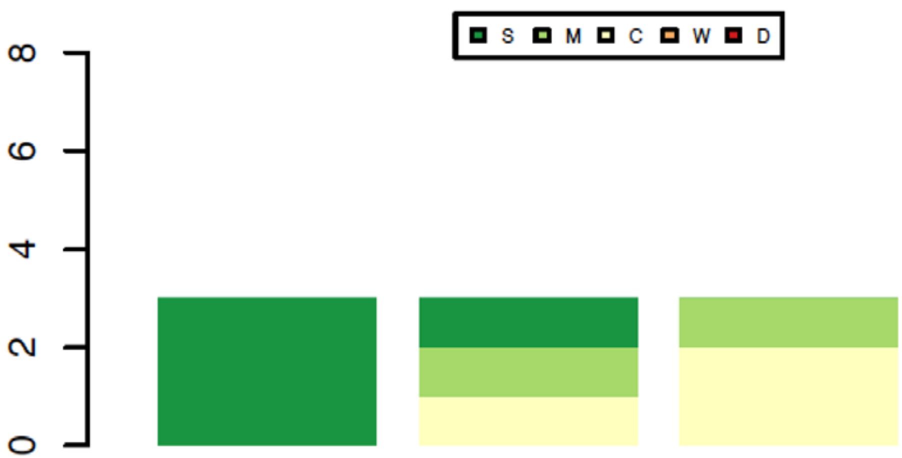
10.6–14.3

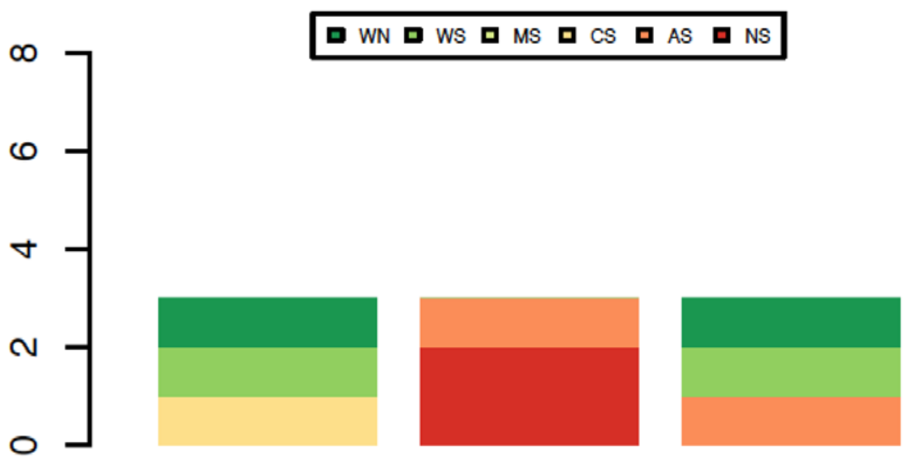
CorrP-AWE
CorrP-AWE-TS
MI-ARACNE-TS



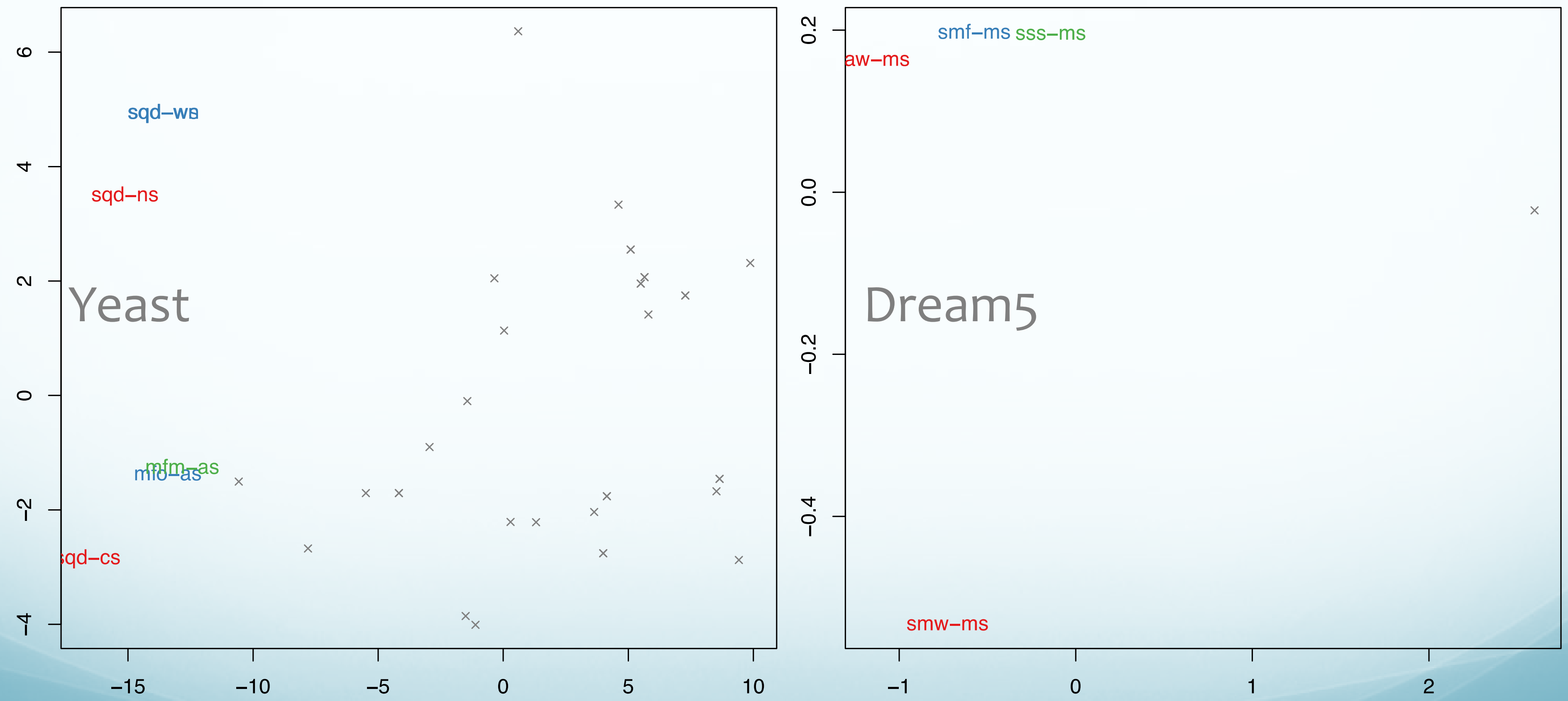
Pareto Fronts Analysis

- Similarity measures (left-hand graph)
  - **Mostly symbolic** (dark green; 4, **all 3 in the first Pareto front**)
  - Some based on **mutual-information** (light green; 3)
  - Others based on **correlation** (yellow; 3)
- Scoring schemes (right-hand graph)
  - **Majority AWE weighting scheme** (dark and light green), **no MRNET**



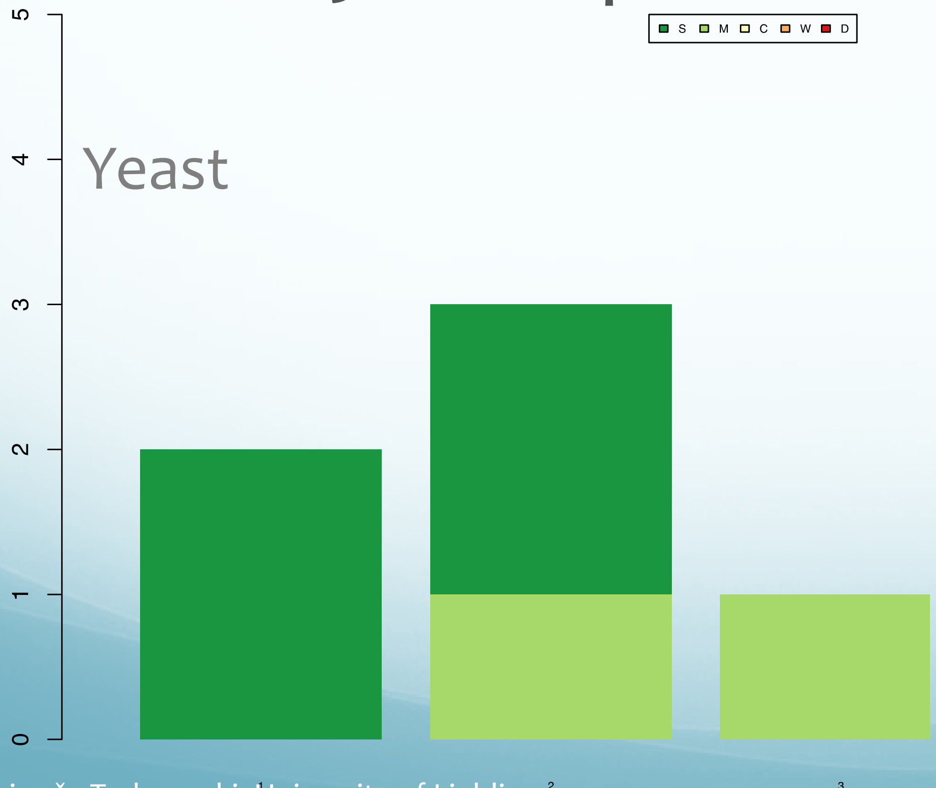


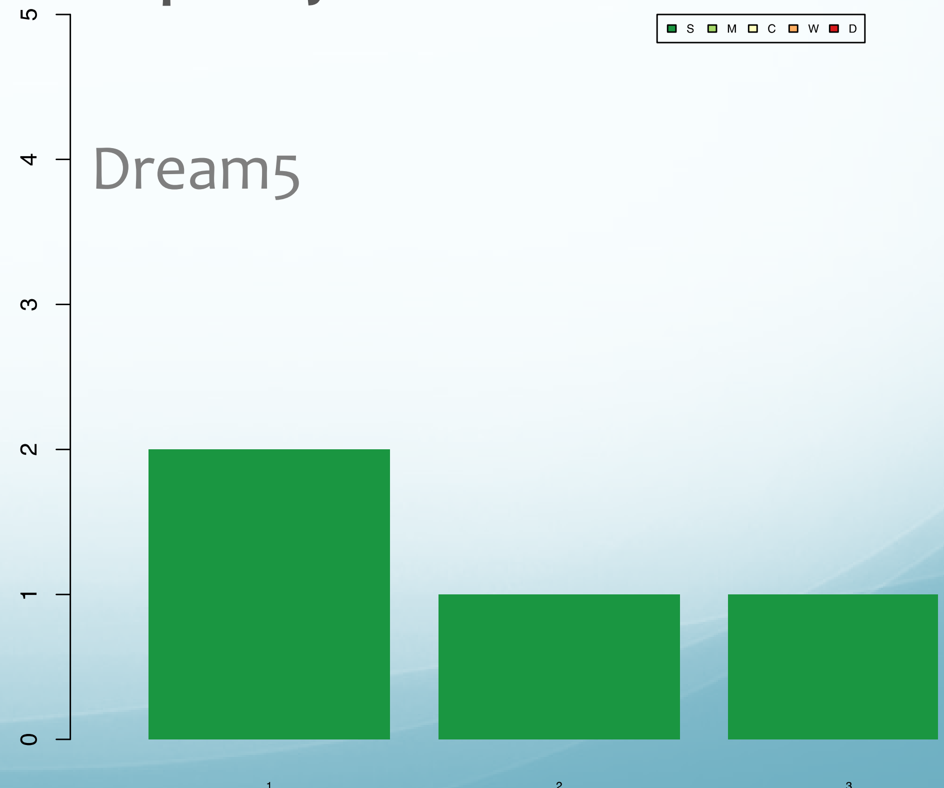
Comparison: Yeast vs. Dream5 (YvD)



YNvDN: Similarity Measures

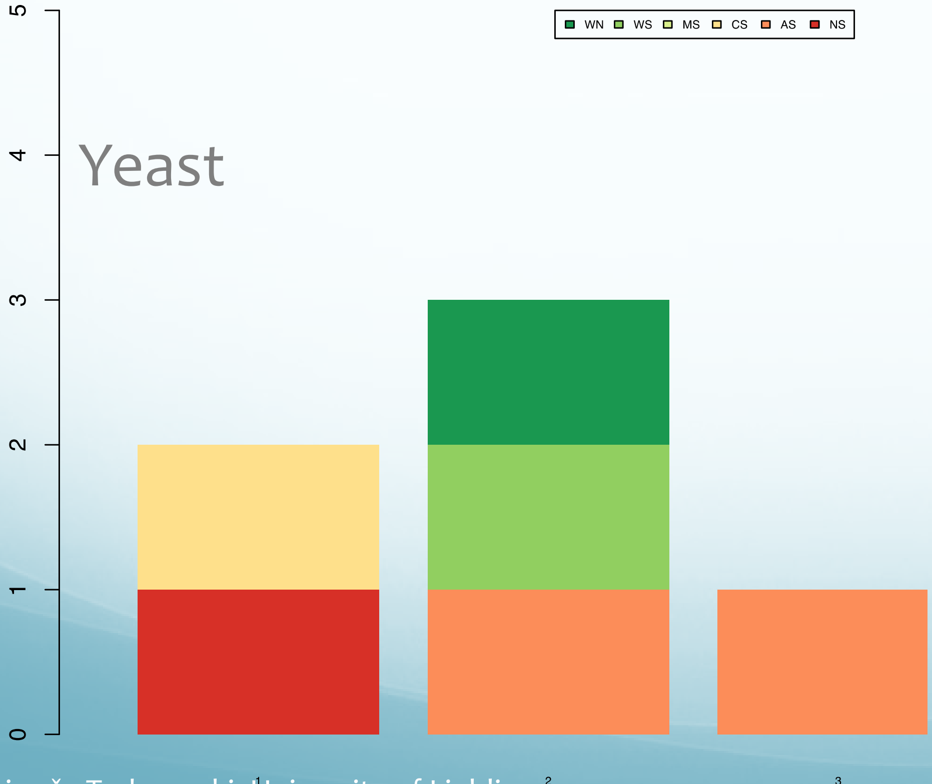
- In both cases: **symbolic measures** (dark green) **perform best**
  - Yeast: also mutual-information (light green) based measures
  - **Yeast:** all **the best symbolic performers** use the **simple QD measure**
  - **Dream5: the best performers** use **complex symbolic measures**

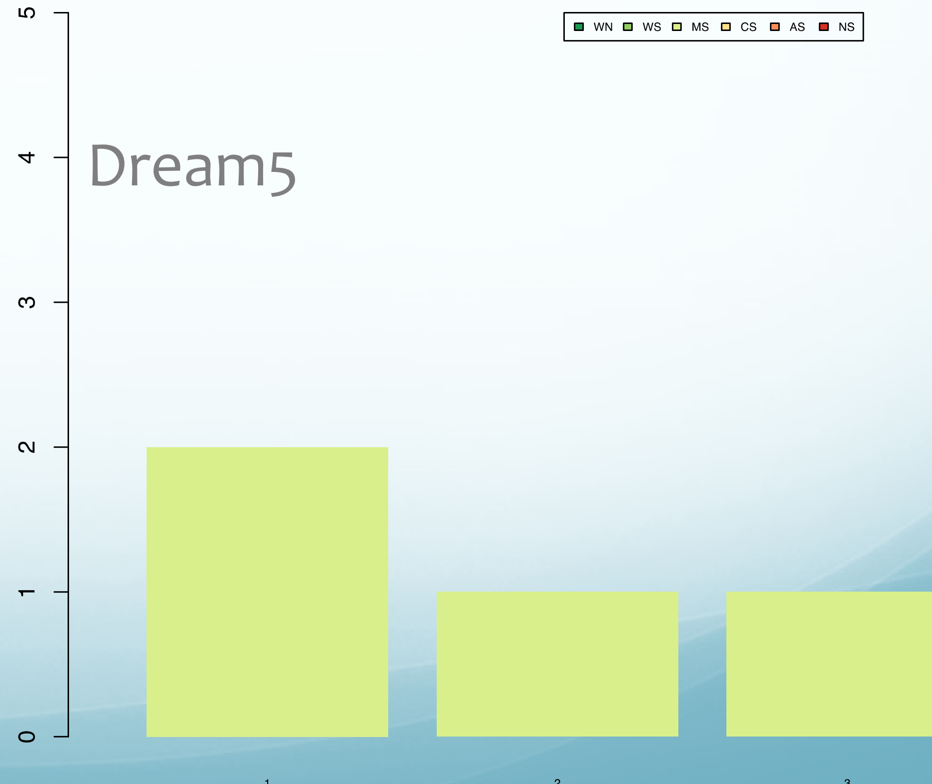




YNvDN: Scoring Schemes

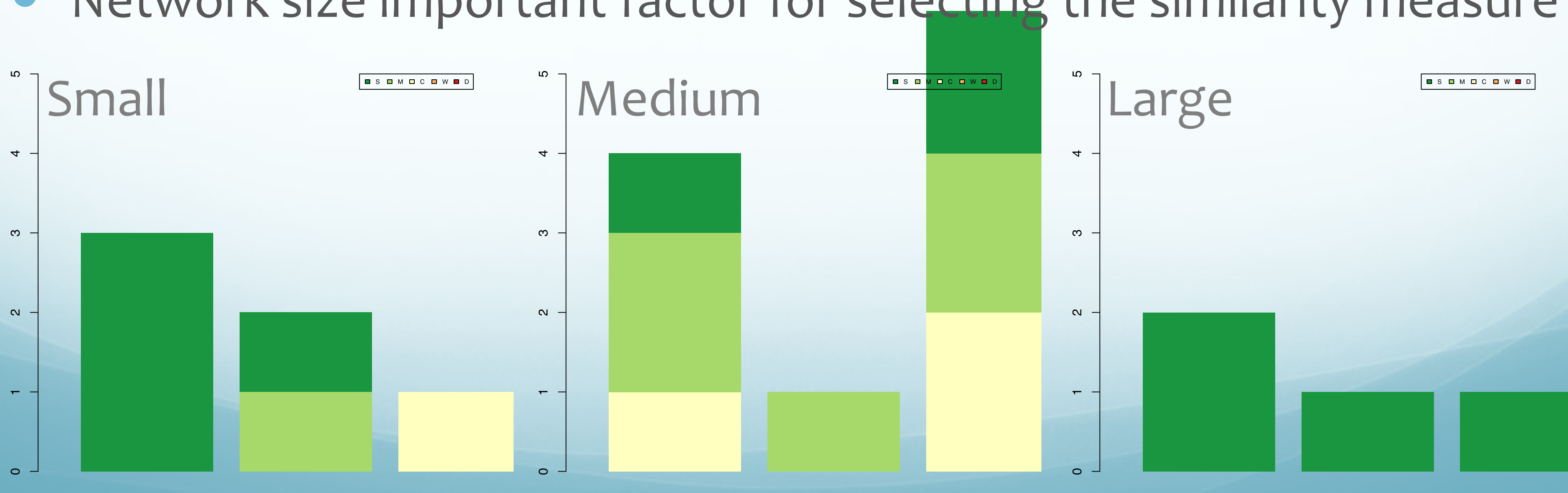
- Difficult to generalize
  - Yeast: five schemes among top performers; **only MRNET missing**
  - Dream5: **MRNET is the only scheme** used by the top performers





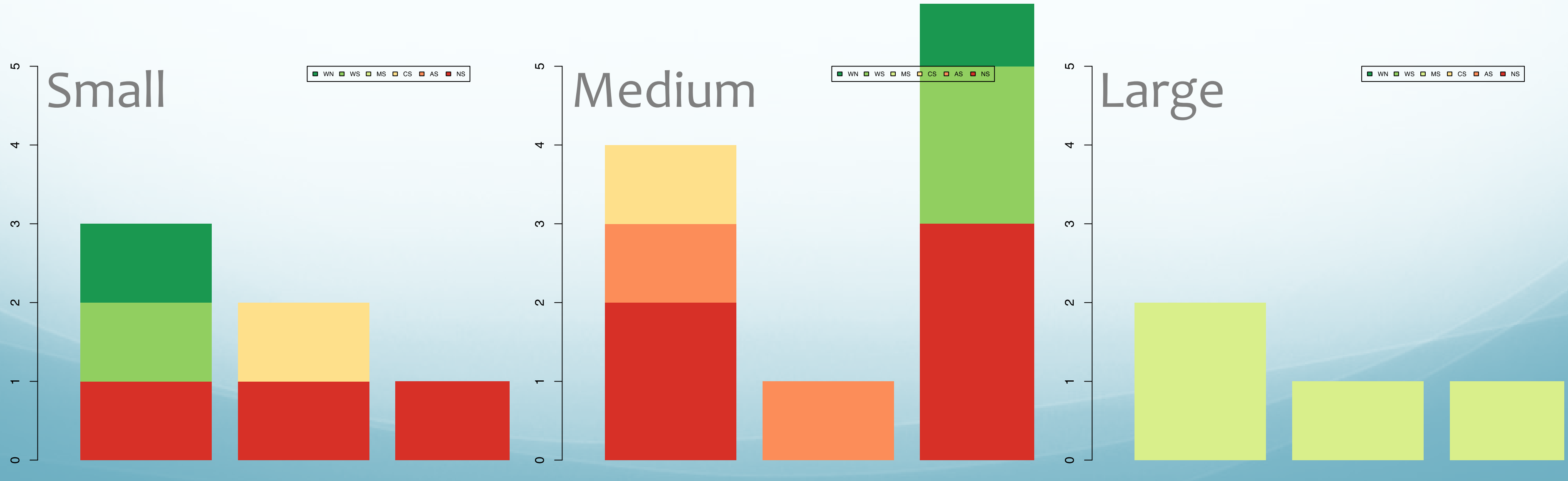
Network Size: Similarity Measures

- Again, symbolic measures prevail; Pearson correl (yellow) for small and medium networks
  - **Small: symbolic** (4; **all simple QD**), mutual info (1) and Pearson (1)
  - **Medium: mutual info** (5), symbolic (3 **simple QD**) and Pearson (3)
  - **Large** networks = Dream5 networks: **complex symbolic**
- Network size important factor for selecting the similarity measure



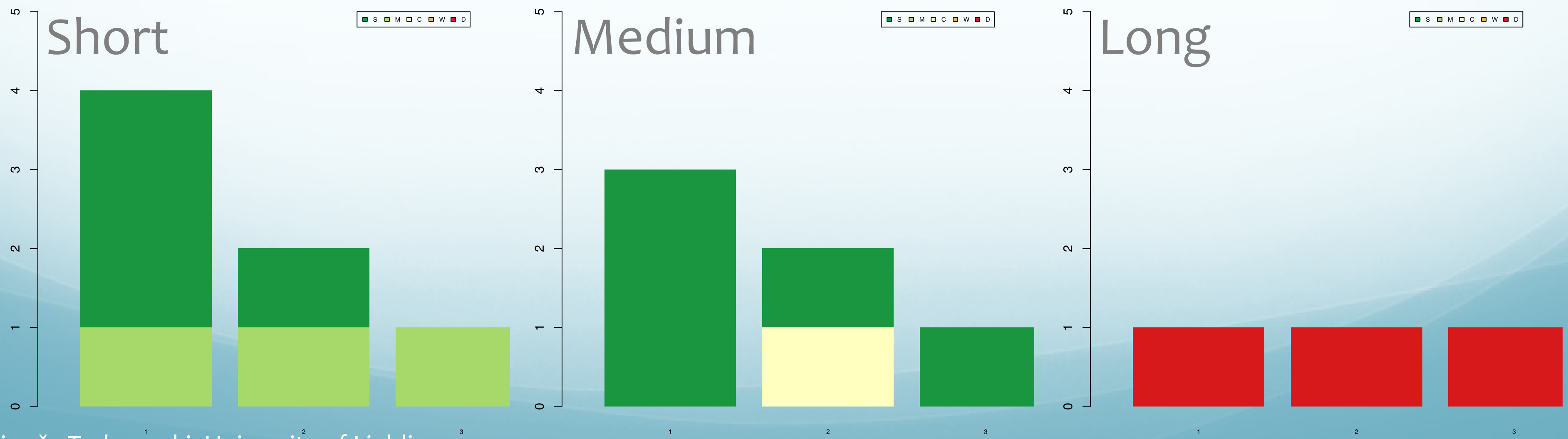
Network Size: Scoring Schemes

- Scoring scheme selection more important for non-small networks
  - Small and medium: 4 and 5 different scoring schemes; **no MRNET**
  - Large networks = Dream5 networks: **MRNET only**
- No obvious relation



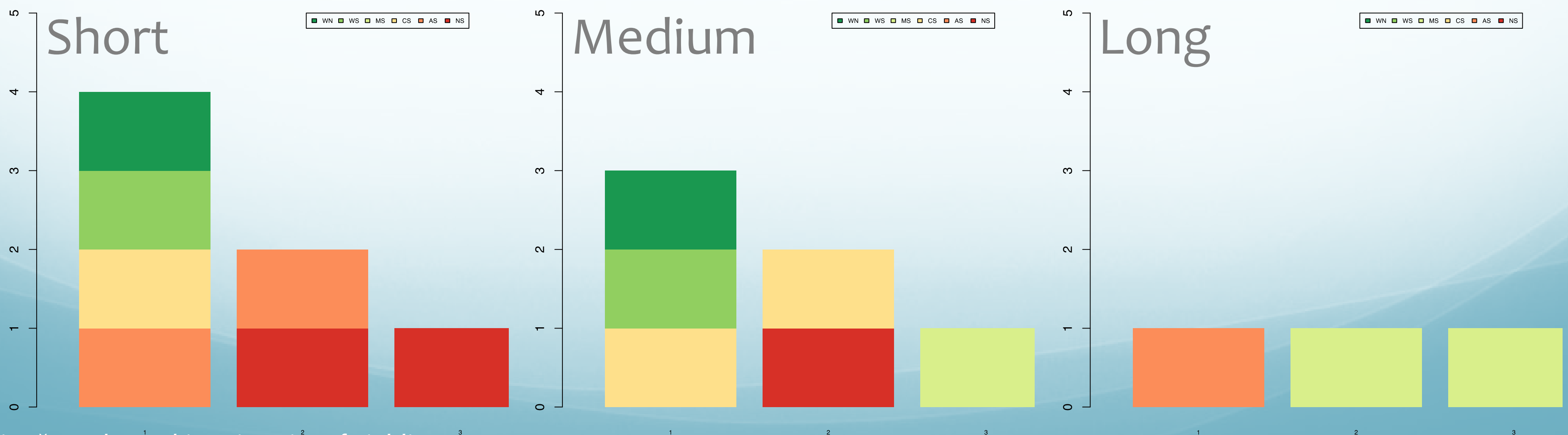
Time Series Length: Similarity Measures

- Time series length important factor for selecting similarity measure
  - **Short: symbolic (QD)** and **mutual-information** based
  - **Medium: symbolic** (mostly QD, **also complex**) and **Pearson corr.**
  - **Long: plain distances** (L10 and Eucledian; red) **perform best**
- **Symbolic measures perform well for not-too-long time series only**



Time Series Length: Scoring Schemes

- TS length not important when selecting the scoring scheme
  - **Small: no MRNET**
  - **Medium: no AWE** (dark orange)
  - Large: AWE and MRNET



Talk Outline

- Introduction and motivation
- Relevance network (RN) approach
  - Similarity measures
  - Scoring schemes
- Empirical comparison of the RN variants
  - Experimental setup: networks, data sets, performance measures
  - Comparison methodology
  - Empirical results: what works where?
- Conclusion and further work

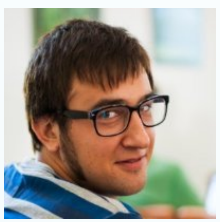
Conclusion: What Works Where?

- Most successful similarities: based on symbolic dynamics
  - The simple qualitative distance measure best overall performer; top performer for small/med networks and short/mid-length time series
  - Complex symbolic measures better for large networks
- Pearson correlation seems to work well for medium networks
  - No other correlations among the top performers
- Distances work well for long time series
  - Distances based on p=10 and Euclidian norm top performers
- Mutual-info top performers for short time series
- No DTW among the top performers

Further Work

- Open issue: similarity measure and scoring scheme combo
  - Which combination work well and which are broken?
- More experiments and benchmarks
  - These might be performed for additional GRN benchmarks
  - Other domains: Social Networks? Collaborative Environments?
- General methodology for comparing methods performance
  - Taking into account multiple perf criteria
  - In contrast with current *average rank diagrams* that are limited to comparing methods wrt one performance criterion
  - Extend the methodology with quantifying and testing the significance of the differences between Pareto fronts

Collaboration and Acknowledgements



Vladimir Kuzmanovski
Jožef Stefan Institute

Ljupčo Todorovski
University of Ljubljana

Sašo Džeroski
Jožef Stefan Institute

MAESTRA
LEARNING FROM MASSIVE, INCOMPLETELY ANNOTATED, AND STRUCTURED DATA