diffuStats: an R package for benchmarking diffusion scores

Sergi Picart-Armada
Dr. Alexandre Perera-Lluna

Bioinformatics and Biomedical Signals Laboratory
Polytechnic University of Catalonia

January 14, 2019

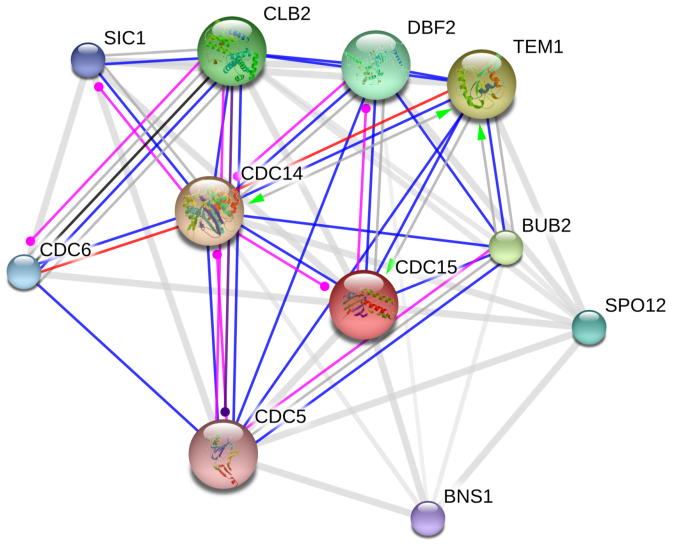
The guilt-by-association principle

Basic version

If two proteins interact with one another, they usually participate in the same, or related, cellular functions [Oliver, 2000].

Generalisations

- Gene-disease association
- Protein label propagation
- Pathway gene prediction
- Metabolomics peak annotation



Small protein-protein interaction network in STRING [Szklarczyk et al., 2015].

Label propagation

Basic idea

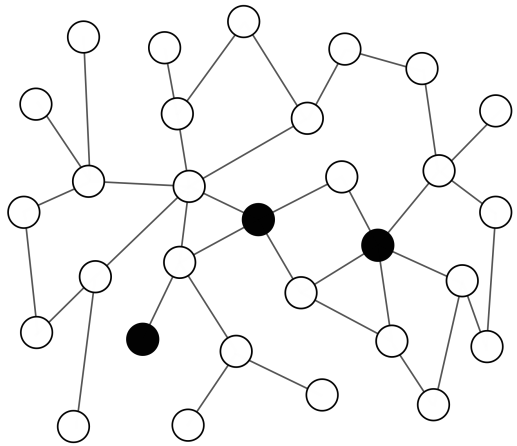
Label propagation spreads the labels from the labelled data to the rest of nodes in semi-supervised learning
[Zoidi et al., 2015].

Diffusion model

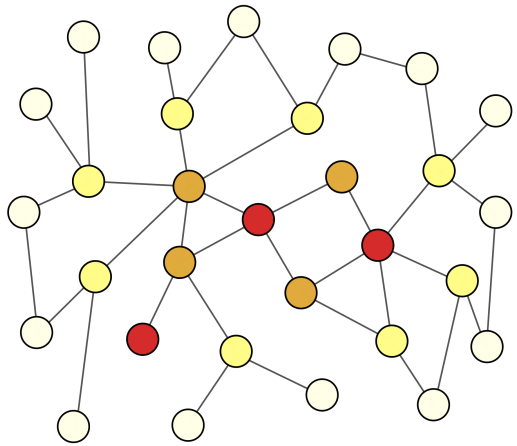
- Basis of several physical models
- Vastly applied in bioinformatics
- Robust to noise
- Intuitive: heat diffusion

Alternatives

- Other graph kernels
- Random walks, e.g. PageRank
- Classical machine learning



Labelled nodes in a network.



Labels have been propagated.

The influence of topology

Node inequivalence issue

The topological properties of the nodes drive the behaviour of their diffusion scores. Some factors:

- Degree – presence of hubs
- Distance to labelled nodes

Idea: background distribution

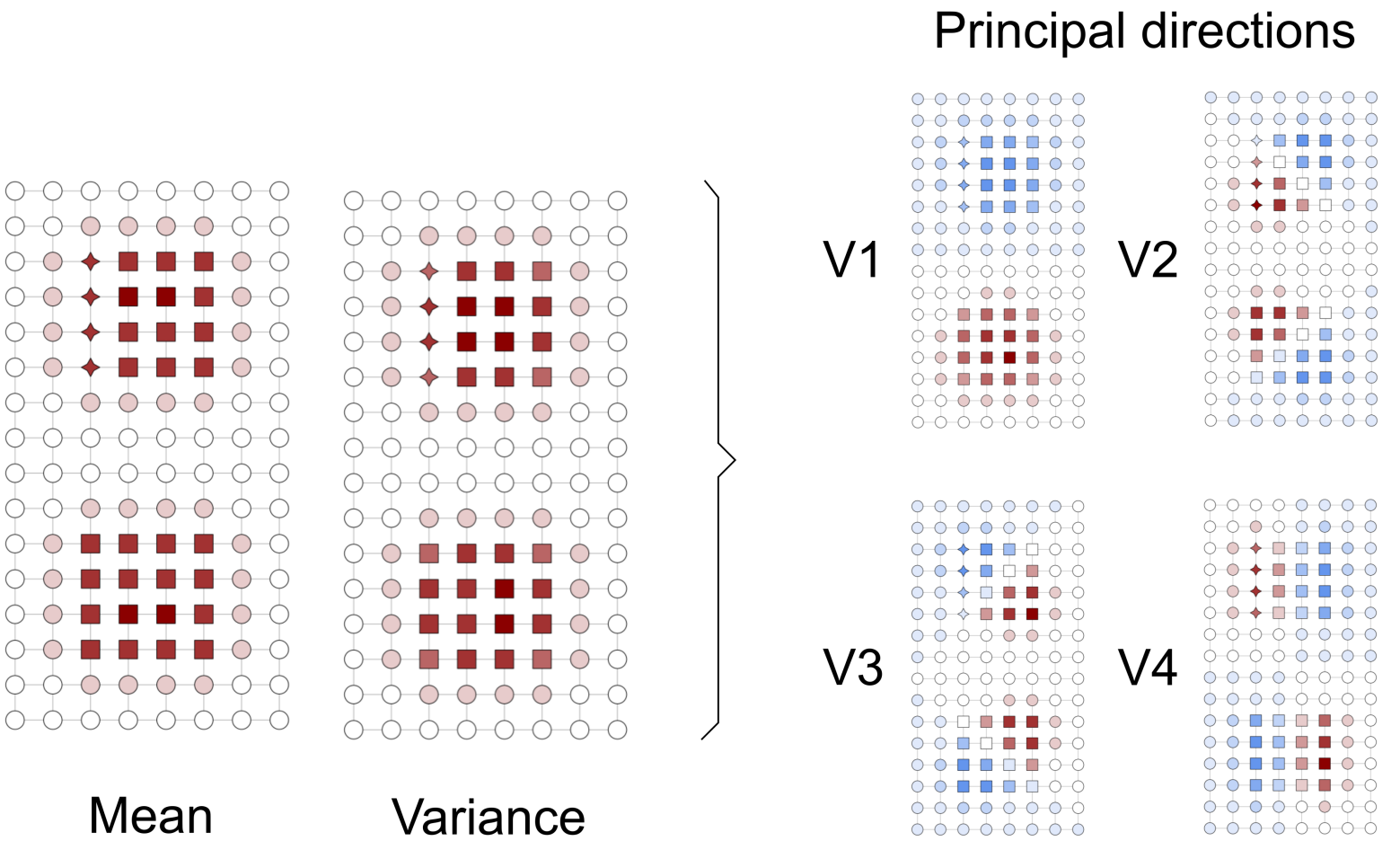
The scores are compared to a null model that attempts to remove the topological bias.
If K is the graph kernel and y the input labels, the diffusion scores

$$f = Ky$$

are compared to their null distribution from a permuted input

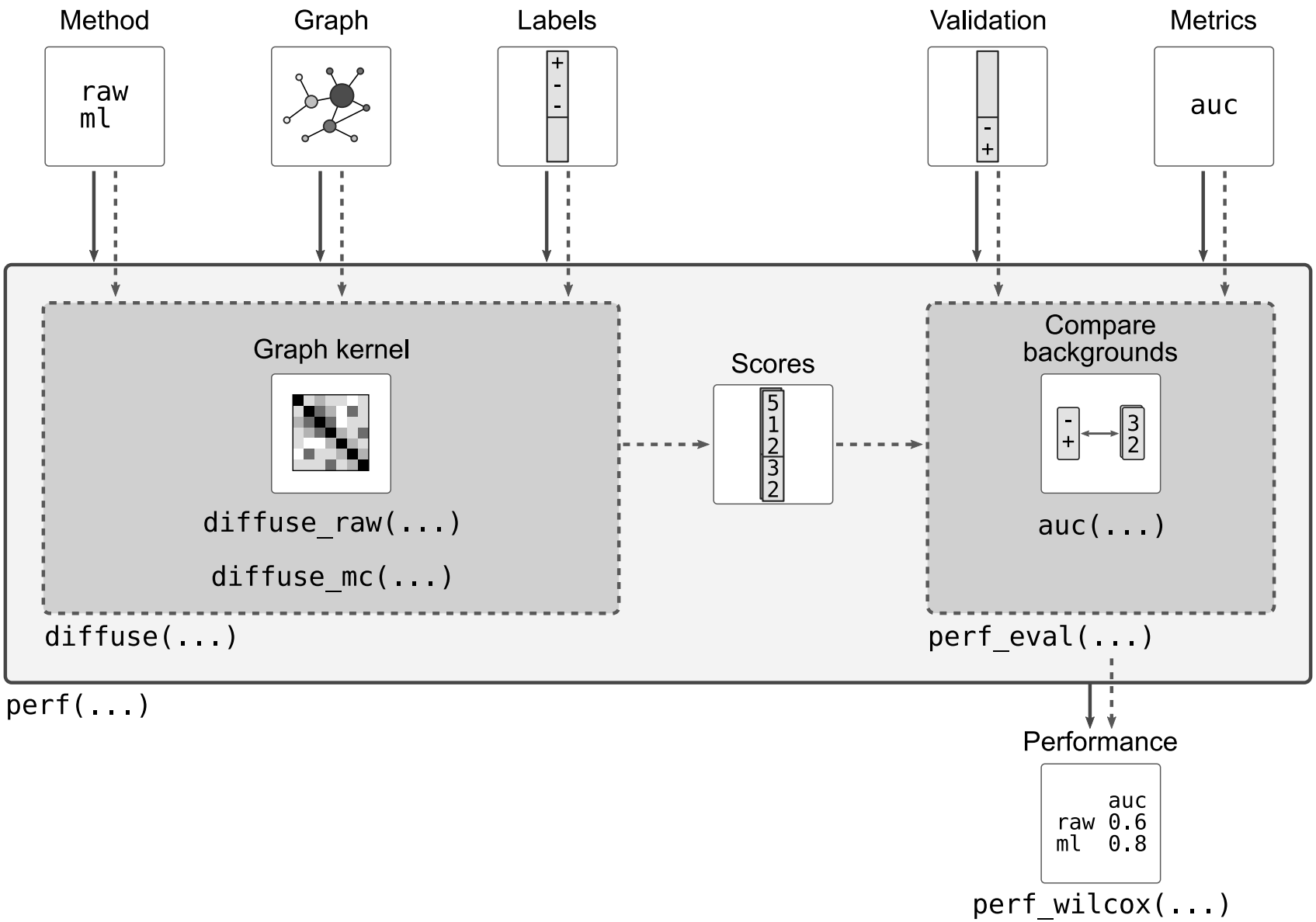
$$f_{null} = K\pi(y)$$

The null distributions



Null mean value and variance of diffusion scores with an input of 4 positive and 28 negative labels. The null variance can be decomposed in its principal directions, aligned with the graph structure.

The R package diffuStats (I)



Workflow for benchmarking diffusion scores using diffuStats [Picart-Armada et al., 2017].

The R package diffuStats (II)

Kernel	Function
Regularised Laplacian	$r(\lambda) = 1 + \sigma^2\lambda$
Diffusion process	$r(\lambda) = \exp(\frac{\sigma^2}{2}\lambda)$
p-Step random walk	$r(\lambda) = (a - \lambda)^{-p}$ with $a \geq 2$, $p \geq 1$
Inverse cosine	$r(\lambda) = (cos(\lambda\frac{\pi}{4}))^{-1}$

Graph kernels available in diffuStats [Picart-Armada et al., 2017].

The R package diffuStats (III)

Score	y^+	y^-	y^u	Normalised	Stochastic	Reference
raw	1	0	0	No	No	Vandin, 2010
ml	1	-1	0	No	No	Tsuda, 2005
gm	1	-1	k	No	No	Mostafavi, 2008
ber_s	1	0	0	No	No	Bersanelli, 2016
ber_p	1	0	0*	Yes	Yes	Bersanelli, 2016
mc	1	0	0*	Yes	Yes	Bersanelli, 2016
z	1	0	0*	Yes	No	Harchaoui, 2013

*unlabelled and negative are inequivalent because permutations affect negatives only. 'gm' has a bias k for unlabelled data, see supplementary

Diffusion scores available in diffuStats [Picart-Armada et al., 2017]. Their differences lie on the codification of the input labels and the statistical normalisation.

Case study

The yeast dataset [Von Mering et al., 2002]

- 2,617 proteins
- 11,855 interactions as edges
- 13 disjoint biological functions

Challenge

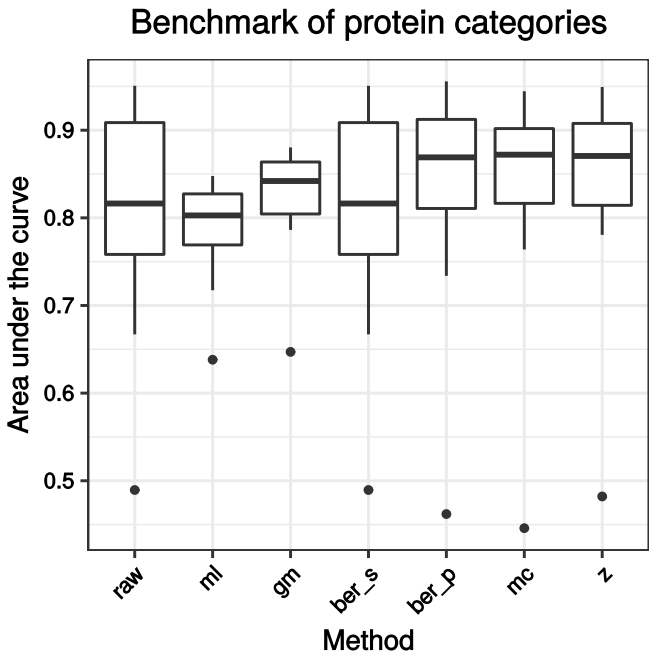
Predict half of the labels from the other half.

Sould we correct that...

- Some nodes tend to have high scores?
- Some nodes show large variances?

Solution

Benchmark all the possibilities!



Benchmark on the diffusion scores in the example dataset from diffuStats [Picart-Armada et al., 2017].

Availability

Bioconductor

diffuStats is actively maintained and available in Bioconductor:
`https://doi.org/doi:10.18129/B9.bioc.diffuStats`

Vignettes with toy and real examples

- Quick start: toy dataset to illustrate the essentials
- Case study: main vignette, thorough description and example

Questions

Status of `diffuStats`:
R package: published [Picart-Armada et al., 2017]
Characterisation: submitting soon
Case study: finding target genes (submitted)

Maintainer:
`sergi.picart@upc.edu`

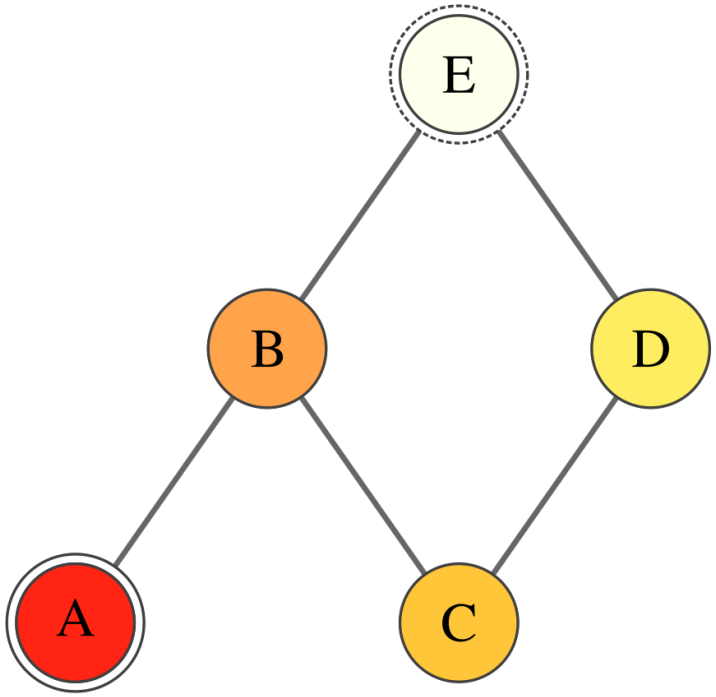
diffuStats: an R package for benchmarking diffusion scores

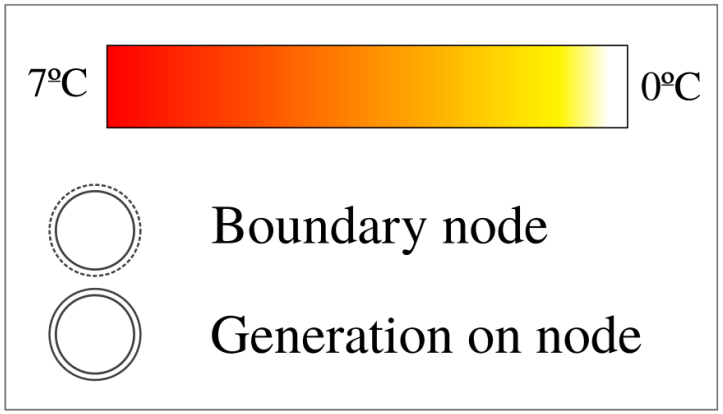
Sergi Picart-Armada
Dr. Alexandre Perera-Lluna

Bioinformatics and Biomedical Signals Laboratory
Polytechnic University of Catalonia

January 14, 2019

Toy example on heat diffusion





$$T = \begin{bmatrix} 7 \\ 3 \\ 2 \\ 1 \end{bmatrix} {}^{\circ}C$$

$$KI = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -3 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} \frac{W}{{}^{\circ}C}$$

$$KC = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \frac{W}{{}^{\circ}C}$$

$$TC = \begin{bmatrix} 0 \end{bmatrix} {}^{\circ}C$$

$$G = \begin{bmatrix} 4 \\ 0 \\ 0 \\ 0 \end{bmatrix} W$$

References I

Oliver, S. (2000).
Proteomics: guilt-by-association goes global.
Nature, 403(6770):601–603.

Picart-Armada, S., Thompson, W. K., Buil, A., and Perera-Lluna, A. (2017).
diffuStats: an R package to compute diffusion-based scores on biological networks.
Bioinformatics, 34(3):533–534.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015).
STRING v10: protein-protein interaction networks, integrated over the tree of life.
Nucleic Acids Research, 43(Database-Issue):447–452.

References II


Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002).
Comparative assessment of large-scale data sets of protein–protein interactions.
Nature, 417(6887):399–403.


Zoidi, O., Fotiadou, E., Nikolaidis, N., and Pitas, I. (2015).
Graph-based label propagation in digital media: A review.
ACM Computing Surveys (CSUR), 47(3):48.