Topological and Geometric Data Reduction **3 May, 2019**

Final Project

Instructor: Yuan Yao *Due: 23:59 Sunday 19 May, 2019*

1 Project Requirement and Datasets

In the below, we list some candidate datasets for your reference. You are also encouraged to work on your own datasets in the final project, upon the approval of the instructor.

1. Pick up ONE (or more if you like) favourite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.
2. Team work: we encourage you to form small team, up to THREE persons per group, to work on the same problem. Each team must submit:

   (a) *ONE report, with a clear remark on each person's contribution.* The report can be in the format of a *technical report within 8 pages*, e.g. NIPS conference style

   `https://nips.cc/Conferences/2016/PaperInformation/StyleFiles`

   or of a *poster*, e.g.

   `https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_poster.pptx`

   (b) *ONE short presentation video within 10 mins*, e.g. in Youtube link. You may submit your presentation slides together with the video link to help understanding.
3. In the report, (1) design or raise your scientific problems (a good problem is often more important than solving it); (2) show your main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance results. Source codes may be submitted through email as a zip file, or as an appendix if it is not large.
4. Submit your report by email or paper version no later than the deadline, to the following address (datascience.hw@gmail.com) with Title: CSIC 5011: Project 2.

Open Peer Review

In this exercise of open peer review, please write down your comments of the *reports rather than of your own team* in the following format. Be considerate and careful with a precise description, avoiding offensive language.

Deadline is 23:59 May 25, 2019. Submit your review in plain text to the email address (datascience.hw@gmail.com) with Title: CSIC 5011: Project 2 Review. Rebuttal is open afterwards.

- Summary of the report.
- Describe the strengths of the report.
- Describe the weaknesses of the report.
- Evaluation on quality of writing (1-5): Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.
- Evaluation on presentation (1-5): Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.
- Evaluation on creativity (1-5): Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations or extensions of existing ideas are also valuable.
- Confidence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

Rebuttal

The rebuttal period starts from now, till 23:59 May 31, 2019. Restrict the number of characters of your rebuttal within **5,000**. Submit your rebuttal in *PLAIN TEXT* or *Word Document* format to the email address (datascience.hw@gmail.com) with Title: CSIC 5011: Project 2 Rebuttal.

The following tips of rebuttal might be helpful for you to follow:

1. The main aim of the rebuttal is to answer any specific questions that the reviewers might have raised, or to clarify any misunderstanding of the technical content of the paper.

2. Keep your rebuttal short, to-the-point, and specific. In our experience, such rebuttals have the maximum impact.

3. Always be polite and professional. Refrain from name calling or rude comments, especially in response to negative reviews.

4. Highlight the changes in your manuscripts had you made a simple revision.

2 Crowdsourced Ranking Data on Allourideas

The following datasets are crowdsourced pairwise ranking from platform Allourideas by Professor Mathew Salganik of Princeton Sociology. You may explore it with HodgeRank etc.

2.1 World College Rankings

The following website hosts the crowdsourcing task on pairwise ranking on 270 universities in the world:

`http://www.allourideas.org/worldcollege`

Up to Nov 26, 2017, the following dataset is collected at github:

`https://github.com/yuany-pku/data/tree/master/allourideas/allourideas_worldcollege`

where you may find

- explanation of data file formats: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/allourideas%20-%20download%20your%20data.pdf`
- 270 universities: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_candidates_2017-11-26T07_14_53Z.csv`
- all valid votings: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_votes_2017-11-26T07_15_02Z.csv`
- all nonvotings: `https://github.com/yuany-pku/data/blob/master/allourideas/allourideas_worldcollege/wikisurvey_colleges_nonvotes_2017-11-26T07_15_30Z.csv`

This dataset has been used for various studies, e.g. Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. False Discovery Rate Control and Statistical Quality Assessment of Annotators in Crowdsourced Ranking, ICML 2016, in `https://arxiv.org/abs/1605.05860v1`. An old dataset cleaned by Prof. Qianqian Xu from CAS can be found at

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/college.csv`

2.2 Human Age Ranking

The following dataset is kindly provided by Qianqian Xu, CAS, for the exploration on class.

The dataset is contained in the following zip file.

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/age.zip`

where you may find

1. `readme.txt`: description of data

2. `Agedata.mat`: data file collected
3. `Groundtruth.mat`: Groundtruth
4. `30 images.zip`: 30 human face images of different ages

The basic problem is to rank the faces according to the ages, using all the information collected so far. A simple sub-problem is rank aggregation of ages from pairwise comparisons. If you are interested, you can try some generalized linear models (Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. HodgeRank on Random Graphs for Subjective Video Quality Assessment. IEEE Transactions on Multimedia, 14(3):844-857, 2012, `https://github.com/yao-lab/yao-lab.github.io/blob/master/reference/TMM12-final.pdf`) on this dataset, such as uniform model, Bradley-Terry model, Thurstone-Mosteller model, and Angular transform model. Compare maximum likelihood estimators and least square ones. The source code of this paper can be found at

`https://github.com/qianqianxu010/TMM2012`

A recent study with wider data is: Qianqian Xu, Jiechao Xiong, Xiaochun Cao, Qingming Huang, Yuan Yao, From Social to Individuals: a Parsimonious Path of Multi-level Models for Crowdsourced Preference Aggregation, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 41(4):844-856, 2019, where the source codes can be downloaded at

`https://github.com/qianqianxu010/TPAMI2018`

3 PageRank and Primary Eigenvectors

The following dataset contains Chinese (mainland) University Weblink during 12/2001-1/2002,

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/univ_cn.mat`

where `rank_cn` is the research ranking of universities in that year, `univ_cn` contains the webpages of universities, and `W_cn` is the link matrix whose $(i, j) - th$ element gives the number of links from university i to j.

1. Compute PageRank with Google's hyperparameter $\alpha = 0.85$;
2. Compute HITS authority and hub ranking;
3. Compare these rankings against the research ranking (you may consider Spearman's ρ and Kendall's τ to compare different rankings);
4. Compute extended PageRank with various hyperparameters $\alpha \in (0, 1)$, investigate its effect on ranking.

For your reference, an implementation of PageRank and HITs can be found at

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/pagerank.m`

The following academic website link collects more countries with university links, for further explorations:

`http://cybermetrics.wlv.ac.uk/database/`

4 Order the faces by Diffusion Map

The following dataset contains 33 faces of the same person ($Y \in \mathbb{R}^{112\times 92\times 33}$) in different angles,

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/face.mat`

You may create a data matrix $X \in \mathbb{R}^{n\times p}$ where $n = 33, p = 112 \times 92 = 10304$
(e.g. `X=reshape(Y,[10304,33])';` in matlab).

1. Explore the Diffusion map, or the second smallest eigenvector of Markov Chains defined on the point cloud data, to order the faces, i.e., let $W_{ij} = \exp(-\|x_i - x_j\|^2/t)$ with $D = \mathrm{diag}(\sum_j W_{ij})$ and define $L = D^{-1}W - I$, clearly $\lambda_0 = 0$ and take the (second) smallest nonzero eigenvalue λ_1 with corresponding eigenvector v_1, sort the faces by values $v_1(i)$, $i = 1, \ldots, n$.
2. Explore the MDS-embedding of the 33 faces on top two eigenvectors: order the faces according to the top 1st eigenvector and visualize your results with figures.
3. Explore the ISOMAP-embedding of the 33 faces on the $k = 5$ nearest neighbor graph and compare it against the MDS results. Note: you may try Tenenbaum's Matlab code
`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/isomapII.m`
4. Explore the LLE-embedding of the 33 faces on the $k = 5$ nearest neighbor graph and compare it against ISOMAP. Note: you may try the following Matlab code
`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lle.m`

You might explore larger datasets with other manifold learning methods, for example the PubFig dataset et al.

`http://www.cs.columbia.edu/CAVE/databases/pubfig/`

5 Transition Paths of Karate Club Network

The following dataset contains a 34-by-34 adjacency matrix A of Zachery's Karate Club Network.

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/karate.mat`

As shown in Figure 1, node 1 represents the coach of the club and node 34 is the owner (president) of the club. The undirected, unweighted edges between nodes represent the affinity relation between club members. The story behind the network is this: the coach would like to raise the instruction fee while the president does not allow this; the conflicts finally result in a fission of the club – the coach leaves the club with his funs and sets up his own club marked in red, and the blue nodes remain in the old club with the president.

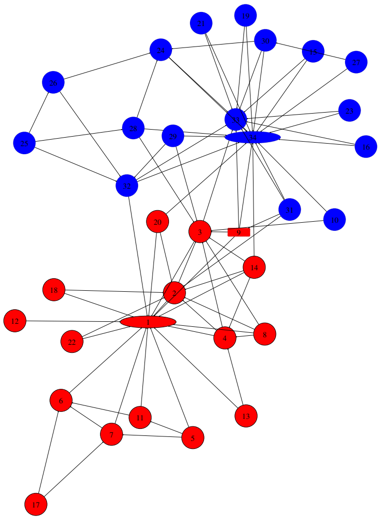

Figure 1: Zachery's Karate Club Network

A. Apply the spectral clustering via the Cheeger vector (the second smallest eigenvector associated with normalized Graph Laplacian) to bipartite the network into two components, and compare it with the ground truth fission above.

B. Perform the following experiment with the transition path analysis.

1. Define a Markov chain according to the network structure, such that from each node a random walker will jump to its neighbors with equal probability, i.e. $P = D^{-1}A$ where $D = \mathrm{diag}(d_i)$ and $d_i = \sum_j A_{ij}$;
2. Compute its stationary distribution $\pi(i) \sim d_i$.
3. Define the source set $V_0 = \{1\}$ and the target set $V_1 = \{34\}$, compute the committor function
$$q(x) = Prob(\text{trajectory starting from } x \text{ hitting } V_1 \text{ before } V_0)$$
by solving the following Dirichlet boundary problem
$$(Lq)(x) = 0, \quad x \in V_u := V - \{1, 34\}, \quad q(1) = 0, \; q(34) = 1.$$
Find those edges which contains one node $q(x) \leq 0.5$ and the other node $q(x) \geq 0.5$. Such edge set defines a cut of the graph.
4. Compute the effective flux on each edge (x, y) by
$$J^+(x, y) = \max(J(x, y) - J(y, x), 0),$$
where
$$J(x, y) = \begin{cases} \pi(x)(1 - q(x))P_{xy}q(y), & x \neq y; \\ 0, & \text{otherwise.} \end{cases}$$

5. Compute the transition flux through each node $x \in V$ by

$$T(x) = \begin{cases} \sum_{y \in V} J^+(x,y), & x \in V_0 \\ \sum_{y \in V} J^+(y,x), & x \in V_1 \\ \sum_{y \in V} J^+(x,y) = \sum_{y \in V} J^+(y,x), & x \in V_u \end{cases}$$

6. Visualize your results by plotting a directed graph, with an arrow on each edge indicating the effective flux direction $J^+(x,y) > 0$, different color marking the cut set of the graph, and if possible edge/node size in proportion to the size of effective/transition flux.

A reference can be seen at:

- Weinan E, Jianfeng Lu, and Yuan Yao. *The Landscape of Complex Networks: Critical Nodes and A Hierarchical Decomposition.* Methods and Applications of Analysis, special issue in honor of Professor Stanley Osher on his 70th birthday, 20(4):383-404, 2013.
  arXiv: `http://arxiv.org/abs/1204.6376`.
  Pdf Link: `https://github.com/yao-lab/yao-lab.github.io/blob/master/reference/ELY.MAA13.pdf`

The following matlab codes implement the transition path analysis and reproduce the results in the paper above:

`https://github.com/yao-lab/yao-lab.github.io/blob/master/data/karate_tpt.m`

You may explore more networks, including the two other examples in the paper above:

LAO-binding network (provided by Xuhui Huang): `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lao54.mat`

Les Miserables social network: `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lesmis.mat`, or `https://github.com/yao-lab/yao-lab.github.io/blob/master/data/lesmis.txt`

6 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

`https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing`

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

`https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit`

contains the labels of such paintings, which are

1 Maybe Raphael - Disputed

2	Raphael
3	Raphael
4	Raphael
5	Raphael
6	Raphael
7	Maybe Raphael - Disputed
8	Raphael
9	Raphael
10	Maybe Raphael - Disputed
11	Not Raphael
12	Not Raphael
13	Not Raphael
14	Not Raphael
15	Not Raphael
16	Not Raphael
17	Not Raphael
18	Not Raphael
19	Not Raphael
20	My Drawing (Raphael?)
21	Raphael
22	Raphael
23	Maybe Raphael - Disputed
24	Raphael
25	Maybe Raphael - Disputed
26	Maybe Raphael - Disputed
27	Raphael
28	Raphael

Can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? The following student poster report seems a good exploration

`https://yao-lab.github.io/2015.fall.pku/poster/Raphael_LI%2CYue_1300010601.pdf`

The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

`http://dx.doi.org/10.1016/j.acha.2015.11.005`

Datasets from Project 1

All the following datasets have been given in project 1. But you may explore these data with various new techniques learned in class, e.g. Robust PCA, Sparse PCA, Robust PCA by GAN (`https://github.com/zhuwzh/Robust-GAN-Scatter`), Manifold Learning, Topological Data Analysis (Mapper, Persistent Homology, etc.).

7 Finance Data

The following data contains 1258-by-452 matrix with closed prices of 452 stocks in SNP'500 for workdays in 4 years.

`https://yao-lab.github.io/data/snp452-data.mat`

or in R:

`https://yao-lab.github.io/data/snp500.Rda`

8 Animal Sleeping Data

The following data contains animal sleeping hours together with other features:

`https://yao-lab.github.io/data/sleep1.csv`

9 US Crime Data

The following data contains crime rates in 59 US cities during 1970-1992:

`https://yao-lab.github.io/data/crime.zip`

Some students in previous classes study crime prediction in comparison with MLE and James-Stein, for example, see https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_slides.pptx

10 NIPS paper datasets

NIPS is one of the major machine learning conferences. The following datasets collect NIPS papers:

10.1 NIPS papers (1987-2016)

The following website:

https://www.kaggle.com/benhamner/nips-papers

collects titles, authors, abstracts, and extracted text for all NIPS papers during 1987-2016. In particular the file `paper_authors.csv` contains a sparse matrix of paper coauthors.

10.2 NIPS words (1987-2015)

The following website:

https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015

collects the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015. The dataset is in the form of a 11463 x 5812 matrix of word counts, containing 11463 words and 5811 NIPS conference papers (the first column contains the list of words). Each column contains the number of times each word appears in the corresponding document. The names of the columns give information about each document and its timestamp in the following format: `Xyear_paperID`.

11 Jiashun Jin's data on Coauthorship and Citation Networks for Statisticians

Thanks to Prof. Jiashun Jin at CMU, who provides his collection of citation and coauthor data for statisticians. The data set covers all papers between 2003 and the first quarter of 2012 from the Annals of Statistics, Journal of the American Statistical Association, Biometrika and Journal of the Royal Statistical Society Series B. The paper corrections and errata are not included. There are 3607 authors and 3248 papers in total. The zipped data file (14M) can be found at

https://yao-lab.github.io/data/jiashun/Jiashun.zip

with an explanation file

https://yao-lab.github.io/data/jiashun/ReadMe.txt

With the aid of Mr. LI, Xiao, a subset consisting 35 COPSS award winners (`https://en.wikipedia.org/wiki/COPSS_Presidents%27_Award`) up to 2015, is contained in the following file

`https://yao-lab.github.io/data/copss.txt`

An example was given in the following article, A Tutorial of Libra: R Package of Linearized Bregman Algorithms in High Dimensional Statistics, downloaded at

`https://arxiv.org/abs/1604.05910`

with the associated R package Libra:

`https://cran.r-project.org/web/packages/Libra/index.html`

The citation of this dataset is: *P. Ji and J. Jin. Coauthorship and citation networks for statisticians. Ann. Appl. Stat. Volume 10, Number 4 (2016), 1779-1812*, (`http://projecteuclid.org/current/euclid.aoas`)

12 Co-appearance data in novels: Dream of Red Mansion and Journey to the West

A 374-by-475 binary matrix of character-event can be found at the course website, in .XLS, .CSV, .RData, and .MAT formats. For example the RData format is found at

`https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/dream.RData`

with a readme file:

`https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/dream.Rd`

as well as the .txt file which is readable by R command `read.table()`,

`https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/HongLouMeng374.txt`

`https://github.com/yuany-pku/dream-of-the-red-chamber/blob/master/README.md`

Thanks to Ms. WAN, Mengting, who helps clean the data and kindly shares her BS thesis for your reference

`https://yao-lab.github.io/reference/WANMengTing2013_HLM.pdf`

Moreover you may find a similar matrix of 302-by-408 for the Journey to the West (by Chen-En Wu) at:

`https://github.com/yuany-pku/journey-to-the-west`

with R data format:

`https://github.com/yuany-pku/journey-to-the-west/blob/master/west.RData`

and Excel format:

`https://github.com/yuany-pku/journey-to-the-west/blob/master/xiyouji.xls`

13 SNPs Data

This dataset contains a data matrix $X \in \mathbb{R}^{p \times n}$ of about $n = 650,000$ columns of SNPs (Single Nucleid Polymorphisms) and $p = 1064$ rows of peoples around the world. Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'), and some missing values marked by 9.

`https://www.dropbox.com/l/scl/AADN8OpaNFy1yB5gyYzNVOfkZGj9SiVDlZo`

which is big (151MB in zip and 2GB original txt). Moreover, the following file contains the region where each people comes from, as well as two variables `ind1` and`ind2` such that $X(\texttt{ind1},\texttt{ind2})$ removes all missing values.

`https://yao-lab.github.io/data/HGDP_region.mat`

More detailed information about these persons in the dataset can be also found at

`https://yao-lab.github.io/data/HGDPid_populations_ALL.xls`

Some results by PCA can be found in the following paper, Supplementary Information.

`http://www.sciencemag.org/content/319/5866/1100.abstract`

14 Protein Folding

Consider the 3D structure reconstruction based on incomplete MDS with uncertainty. Data file:

`http://yao-lab.github.io/data/protein3D.zip`

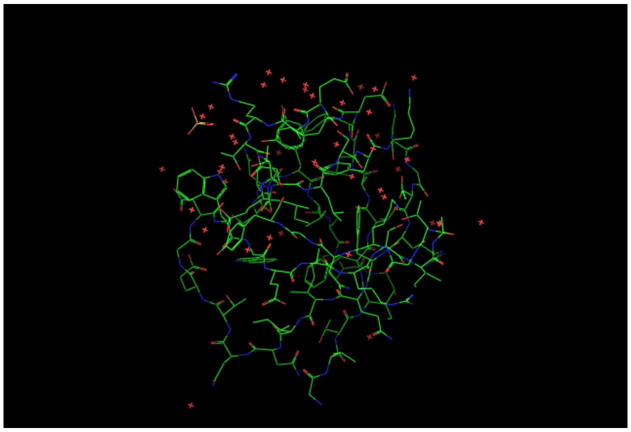

Figure 2: 3D graphs of file PF00018_2HDA.pdf (YES_HUMAN/97-144, PDB 2HDA)

In the file, you will find 3D coordinates for the following three protein families:

PF00013 (PCBP1_HUMAN/281-343, PDB 1WVN),

PF00018 (YES_HUMAN/97-144, PDB 2HDA), and

PF00254 (O45418_CAEEL/24-118, PDB 1R9H).

For example, the file `PF00018_2HDA.pdb` contains the 3D coordinates of alpha-carbons for a particular amino acid sequence in the family, YES_HUMAN/97-144, read as

`VALYDYEARTTEDLSFKKGERFQIINNTEGDWWEARSIATGKNGYIPS`

where the first line in the file is

97 V 0.967 18.470 4.342

Here

- '97': start position 97 in the sequence
- 'V': first character in the sequence
- $[x, y, z]$: 3D coordinates in unit $\mathring{A}$.

Figure 2 gives a 3D representation of its structure.

Given the 3D coordinates of the amino acids in the sequence, one can computer pairwise distance between amino acids, $[d_{ij}]^{l\times l}$ where l is the sequence length. A *contact map* is defined to be a graph $G_\theta = (V, E)$ consisting l vertices for amino acids such that and edge $(i, j) \in E$ if $d_{ij} \leq \theta$, where the threshold is typically $\theta = 5\mathring{A}$ or $8\mathring{A}$ here.

Can you recover the 3D structure of such proteins, up to an Euclidean transformation (rotation and translation), given noisy pairwise distances restricted on the contact map graph G_θ, i.e. given noisy pairwise distances between vertex pairs whose true distances are no more than θ? Design a noise model (e.g. Gaussian or uniformly bounded) for your experiments.

When $\theta = \infty$ without noise, classical MDS will work; but for a finite θ with noisy measurements, SDP approach can be useful. You may try the matlab package SNLSDP by Kim-Chuan Toh, Pratik Biswas, and Yinyu Ye, downladable at `http://www.math.nus.edu.sg/~mattohkc/SNLSDP.html`.