International Journal of Bifurcation and Chaos

RELIABILITY OF OPTIMAL LINEAR PROJECTION OF GROWING SCALE-FREE NETWORKS

PAU EROLA
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
pau.erola@urv.cat

JAVIER BORGE-HOLTHOEFER
Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza
Mariano Esquillor s/n, 50018 Zaragoza, Spain
borge.holthoefer@gmail.com

SERGIO GOMEZ
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
sergio.gomez@urv.cat

ALEX ARENAS
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
alexandre.arenas@urv.cat

Received (to be inserted by publisher)

Singular Value Decomposition (SVD) is a technique based on linear projection theory, which has been frequently used for data analysis. It constitutes an optimal (in the sense of least squares) decomposition of a matrix in the most relevant directions of the data variance. Usually, this information is used to reduce the dimensionality of the data set in a few principal projection directions, this is called Truncated Singular Value Decomposition (TSVD). In situations where the data is continuously changing the projection might become obsolete. Since the change rate of data can be fast, it is an interesting question whether the TSVD projection of the initial data is reliable. In the case of complex networks, this scenario is particularly important when considering network growth. Here we study the reliability of the TSVD projection of growing scale free networks, monitoring its evolution at global and local scales.

Keywords: truncated singular value decomposition; stability; evolving graph

1. Introduction

There exists a vast literature that acknowledges Singular Value Decomposition as a valuable tool for information extraction from matrix-shaped data. This approach and its truncated variant have been extraordinarily successful in many applications [Golub & Van Loan, 1996], in particular for the analysis of relationships between a set of documents and the words they contain. In this case, the decomposition yields

information about word-word, word-document and document-document semantic associations; the technique is known as latent semantic indexing [Berry *et al.*, 1995] (LSI) or latent semantic analysis [Landauer & Dumais, 1997] (LSA). In the field of complex networks, we recently introduced SVD as a useful tool to scrutinize the modular structure in networks [Arenas *et al.*, 2010].

Remarkably, a common characteristic of these applications is their dynamic nature. In order to attain successful information retrieval, for instance in a query, LSI or LSA must rely on the fact that SVD of textual resources is always up to date. Unfortunately, databases rarely stay the same. Addition and/or removal of information is constant, meaning that catalogs and indexes quickly become obsolete or incomplete. Turning to networks, the question is equally pertinent: both natural and artificial networks are dynamic, in the sense that they change through time (and so do their modular structures). Paradigmatic examples of this fact are the Internet, the World Wide Web or knowledge databases like Wikipedia: all of them have been object of study from a graph-theoretical point of view [Pastor-Satorras & Vespignani, 2004; Capocci *et al.*, 2006; Zlatić *et al.*, 2006]. Given this realistic scenario, a major question arises, namely, for how long TSVD stands as a reliable projection of evolving data.

In this paper we study the stability of TSVD as applied on changing networks. In particular, we want to quantify the differences between successive TSVD projections of evolving networks. To this end we devise a set of measures of global and local reliability, and apply them to a classical model of network growth, the Barabási-Albert's (BA) scale-free network [Barabási & Albert, 1999]. The BA model consists in a random network whose formation is driven by: growth, the network starts with a small number of nodes, and a new one is added at each time step; and preferential attachment, the probability of a new node i linking to a previously existing node j is proportional to the current degree of node j. This mechanism yields networks with scale-free degree distributions $P(k) = k^{-3}$.

This work is partially motivated by the application of TSVD to analyze the mesoscale of networks and its temporal evolution. In [Arenas *et al.*, 2010], the object of analysis is the *contribution matrix* C, of N nodes to M modules, where a module (or community) is a set of nodes with more connections between them than with the rest of the network. The rows of C correspond to nodes, and the columns to modules. The analysis of this matrix is the focus of our research. The elements $C_{i\alpha}$ are the number of links that node i dedicates to module α, and is obtained as the matrix multiplication between the network's adjacency matrix A and the *partition matrix* S:

$$C_{i\alpha} = \sum_{j=1}^{N} A_{ij} S_{j\alpha} \,, \tag{1}$$

where $S_{j\alpha} = 1$ if node j belongs to module α, and $S_{j\alpha} = 0$ otherwise. Note that certain changes in the topology might not be reflected in the values of C, for example the rewiring of the connections of a node towards other nodes in the same community. In [Arenas *et al.*, 2010] it was supposed that the modules where not overlapping, thus forming a partition, but the same analysis could have been done with overlapping and fuzzy memberships of nodes to modules.

To measure the reliability of the TSVD projection of the contribution matrix, here we will consider the "worst case scenario" where each node belongs to its own community. This case corresponds mathematically to $C = A$. Establishing that TSVD is robust to change in these circumstances will settle the fact that TSVD is robust to change on a coarse-grained structure.

2. Analysis of networks based on TSVD

Given a rectangular $N \times M$ (real or complex) matrix A, SVD stands for the factorization into the product of three other matrices,

$$A = U \Sigma V^{\dagger} \,, \tag{2}$$

where U is an unitary N-by-N matrix (left singular vectors), and describes the original row entities as vectors of derived orthogonal factor values; Σ, the singular values, is a diagonal N-by-M matrix containing scaling values; and $V^{\dagger}$ denotes the conjugate transpose of V, an M-by-M unitary matrix, which describes the original column entities in the same way as U.

A practical use of SVD is dimensional reduction approximation via truncation, TSVD. It consists in keeping only some of the largest singular values to produce a least squares optimal, lower rank order approximation. For example, severe dimensional reduction is a condition for success in machine learning SVD applications [Deerwester *et al.*, 1990; Berry *et al.*, 1995; Landauer *et al.*, 1998].

In the case of a rank $r = 2$ approximation, the first two left and right singular vectors are unique (up to a sign) if the ordered singular values σ_i of the matrix Σ satisfy $\sigma_1 > \sigma_2 > \sigma_3$ [Golub & Van Loan, 1996]. This dimensional reduction is particularly interesting to depict results in a two-dimensional plot for visualization purposes. Here we suppose that $\text{rank}(A) \geq 2$.

The idea we developed in our previous work [Arenas *et al.*, 2010] is to compute the projection of the connectivity of nodes (rows in A) into the space spanned by the first two left singular vectors, we call this the projection space $\mathcal{U}_2$ and we denote the projected vector of the i-th node as $\tilde{n}_i$. More precisely, if U_2 is the matrix formed by the first two columns of U, then $\tilde{n}_i$ is the i-th row of U_2. Given that the transformation is information preserving [Chu & Golub, 2005] up to rank $r = 2$, the map obtained gives an accurate representation of the main characteristics of the original data, visualizable and, in principle, easier to scrutinize. It is important to highlight that this approach has essential differences with classical pattern recognition techniques based on TSVD such as Principal Components Analysis (PCA) or, equivalently, Karhunen-Loeve expansions. Our data (columns of A) can not be independently shifted to mean zero without loosing its original meaning, this restriction prevents the straightforward application of the mentioned techniques.

To interpret correctly the outcome of the TSVD we change to polar coordinates, where for each node i the radius R_i measures the length of its contribution projection vector $\tilde{n}_i$, and θ_i the angle between $\tilde{n}_i$ and the horizontal axis. Large values of R correspond to highly connected nodes, and θ reflects the adjacencies of each node in matrix A. Fig. 1 shows the R–θ planes of an evolving network to get a visual intuition of the map's stability: as the network grows the mapping is distorted. In the following section we develop measures to quantify the effect of the growth on the TSVD projection.

3. Quantifying the reliability of TSVD on growing networks

As stated in the introduction, the goal of this research is to test how TSVD projection, at rank $r = 2$, changes by computing it at different stages of the evolution of BA scale-free networks. This implies that TSVD will be computed on an initial network of size N_0, and then re-computed for successive node additions up to a final size $N_f = 2N_0$. To quantify the effect of growth on TSVD projection, we devise two levels of study: global and local. We will define measures based on the concept of absolute and relative distances between nodes, to this end we will work in the metric space $\mathcal{U}_2$.

3.1. *Global measure*

We propose a global quantity that indicates the amount of change in the position of nodes in the map obtained by TSVD. In the sequence of computed TSVD projections, the nodes' coordinates in $\mathcal{U}_2$ space change. This can be quantified by the difference of vectors $\tilde{n}_i$ between the initial and evolved network projection.

In Fig. 2 we plot the projection of the growing network presented in Fig. 1 on the space $\mathcal{U}_2$. We fix our attention in two time-shots of the evolution corresponding to growths of 30% and 80%. We compute the differences between positions of the same nodes at different stages (z) as $\tilde{v}_i = \tilde{n}_i^0 - \tilde{n}_i^z$, producing a field map that accounts for the changes. This field map is shown in the insets of Fig.2. When we have a 80% increase of the initial size, the vectors $\tilde{v}_i$ are longer than in the 30% increment, which evidences a larger variability, i.e. a progressive degradation in the TSVD reliability.

The *global error* (E_{global}) measure we propose to assess successive changes of rank r TSVD projection compared to the initial data is computed by the expression

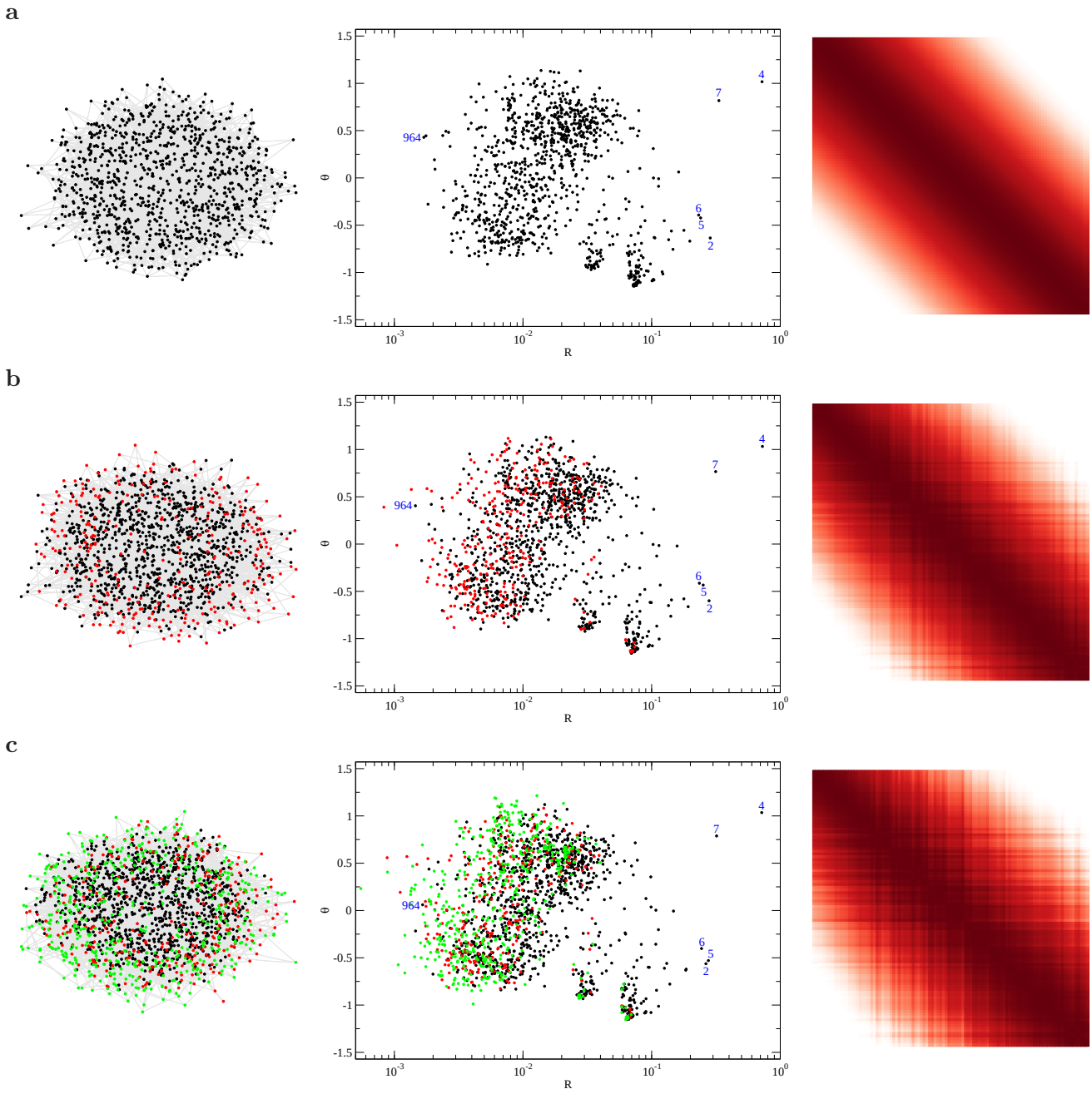

Fig. 1. Three snapshots of a growing network (left side), their corresponding projection on R–θ plane (center) and the θ-overlapping matrices (right side; see [Arenas *et al.*, 2010] for details). For the sake of clarity, the initial set of nodes **(a)** $N = 1000$ are drawn in black; the second snapshot **(b)** represents a growth of a 30% of nodes, $N = 1300$, new nodes are drawn in red. Finally **(c)** represents a network with $N = 1800$, last arrived nodes are depicted in green. Some nodes from the initial set have been highlighted (2, 4, 5, 6, 7, 964) in the R–θ plane, to get a visual intuition of the map's stability. Note that nodes with a high value of R (2, 4, 5, 6, 7) remain almost unchanged throughout the topology's growth; whereas node 964 undergoes much change from an absolute point of view. The rightmost matrices illustrate the amount of change of nodes with respect to their θ angles: as nodes are added in the structure the cosine overlaps between them increasingly distorts the original figure.

$$E_{\text{global}} = \frac{\sum_{i=1}^{N_0} \sum_{j=1}^{r} |U_{ij}^{z} - U_{ij}^{0}|}{\sum_{i=1}^{N_0} \sum_{j=1}^{r} |U_{ij}^{0}|}, \tag{3}$$

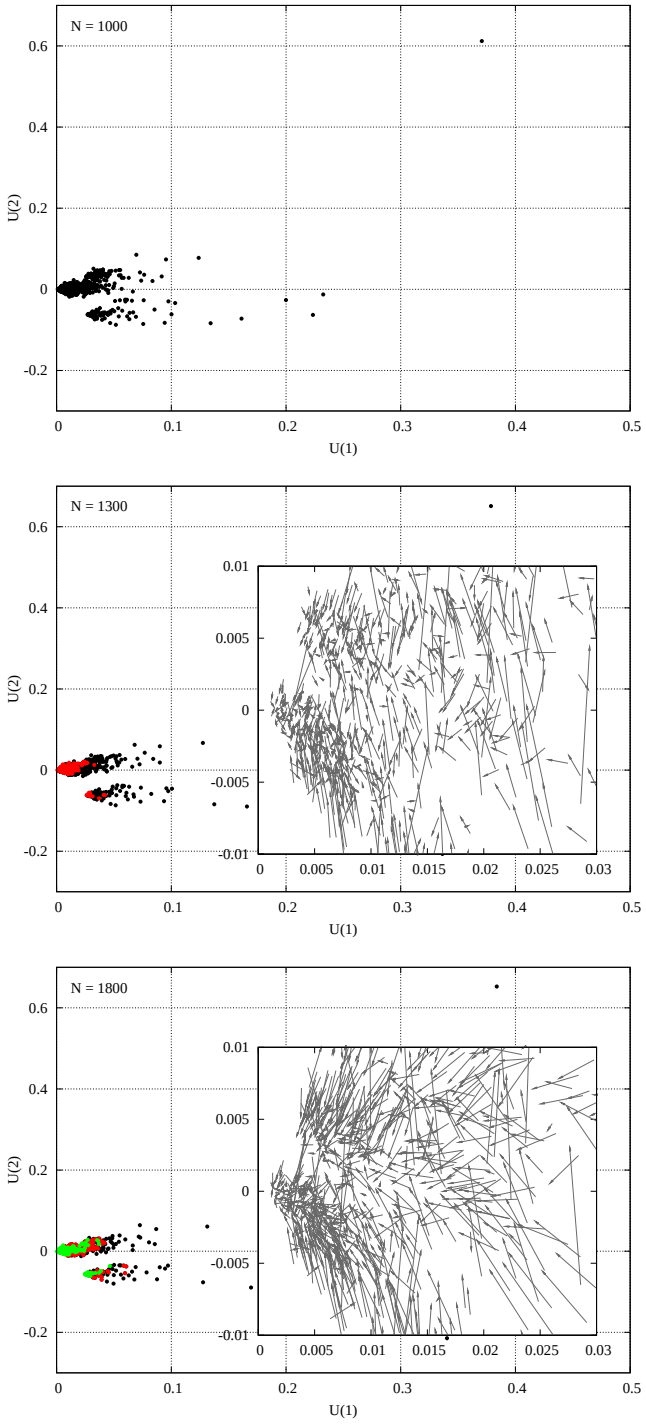

Fig. 2. Projection in the $\mathcal{U}_2$ space (U(1) and U(2) are respectively the first and the second components of the vectors $\tilde{n}_i$) of the evolving network presented in Fig. 1. The insets for $N = 1300$ and $N = 1800$ trace the vectors $\tilde{v}_i$ between the projected coordinates of each node on the grown network and the original coordinates on the initial network with $N_0 = 1000$. We use these vectors $\tilde{v}_i$ to quantify the variability of the TSVD. Nodes are colored like in Fig. 1.

where U^0 represents the truncated left singular vectors of the original network with N_0 nodes; and U^z also represents the truncated left singular vectors, but of the grown network with size $N_z > N_0$.

We have applied this global measure to monitor the evolution of the TSVD stability for growing networks with initial sizes $N_0 = 1000$ and $N_0 = 10000$. Fig. 3 shows the percentage of relative error with

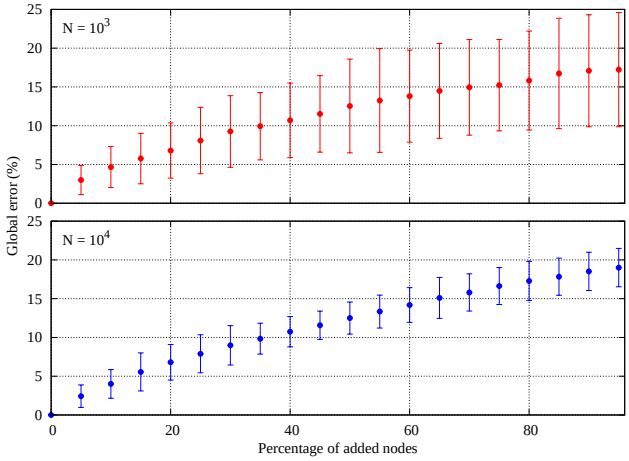

Fig. 3. Global error on two growing networks with initial sizes $N_0 = 1000$ (above) and $N_0 = 10000$ (below). For each network we compute the error by increments of 5% of growth. In both cases, the global error is lower than 10% up to the 40% increment of network size. Each point is the average of 100 simulations.

respect to the original network. In the chart, each successive point represents a 5% of nodes addition. Up to a 40% growth the global error remains below 10%, and doubling the network size the average error still remains below 20%. These results show the reliability of the projection after the growing process.

3.2. *Local measure*

Though informative, the previous global quantity can overlook changes at the microscopic level. The neighborhood of each node in the $\mathcal{U}_2$ plane could undergo changes in the sequence of computed TSVD projections difficult to be revealed by the global measure defined above. Thus, we propose a measure that reflects these local changes using the distances between nodes in a neighborhood. Instead of defining a sharp border for the neighbors of each node, we propose to use a gaussian neighborhood that weights the distances according to a variance σ.

First, we construct the $N \times N$ matrix of distances between any pair of nodes in the network at stage z as:

$$D_{ij}^z = \sqrt{\sum_{k=1}^{r} \left(U_{ik}^z - U_{jk}^z\right)^2}, \tag{4}$$

where U^z represents the truncated left singular vectors of the network. These distances reflect a measure of proximity between nodes, independently on the global positioning in the map. The neighborhood is weighted to prioritize the stability on closer nodes over the distant ones. To this respect, we compute a matrix of weighted distances S^z using a gaussian distribution that establishes a radius of influence as follows:

$$S_{ij}^z = D_{ij}^z e^{-\frac{{D_{ij}^z}^2}{2\left(\sigma R_i^0\right)^2}}, \tag{5}$$

where we have chosen a radius of influence depending on the node. R_i^0 is the module of the projected vector $\tilde{n}_i$ in the initial network, and σ is a constant. This radius of influence proportional to the distance to the origin, emphasizes nodes with larger R which are the most connected ones, see [Arenas *et al.*, 2010]. Using different values of σ in the gaussian function we can tune the size of the neighborhood. Fig. 4 shows, for a network with 1000 nodes, three magnified views of a network projection in $\mathcal{U}_2$ to illustrate the gaussian radius of influence.

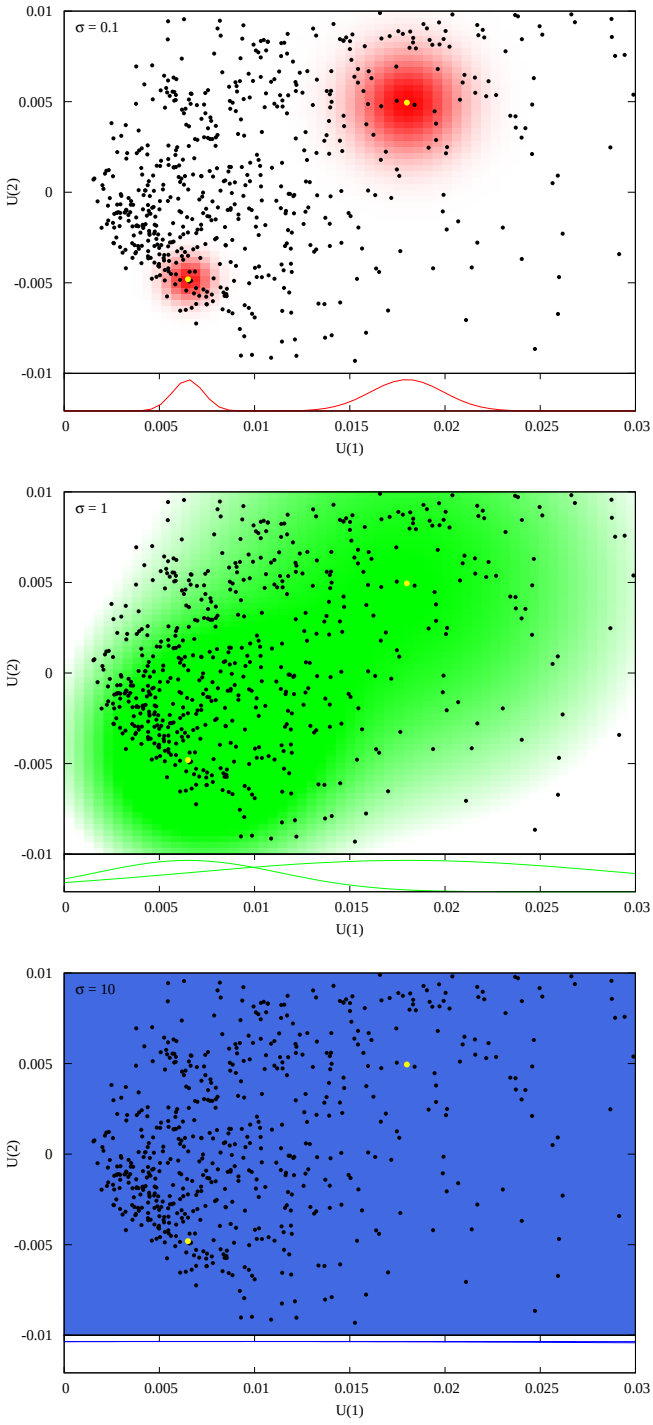

Fig. 4. From top to bottom, we present the radius of influence of $\sigma = 0.1$ (red), $\sigma = 1$ (green) and $\sigma = 10$ (blue) in the projection space $\mathcal{U}_2$. In the bottom of each chart, we have plotted, for the nodes highlighted in yellow, the gaussian curves that we added to matrix D^z to compute the matrix of weighted distances S^z.

Finally, the *local error* (E_{local}) measure of reliability we propose is computed as the *relative error*:

$$E_{\text{local}} = \frac{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} |S^z_{ij} - S^0_{ij}|}{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} |S^0_{ij}|}, \tag{6}$$

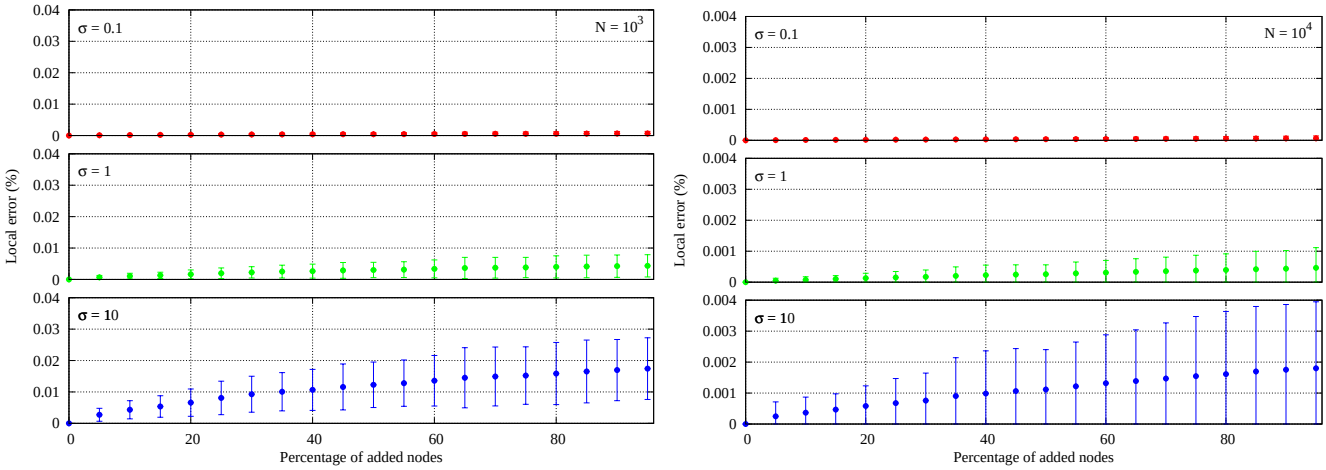

Fig. 5. Local error on two growing networks with initial size $N_0 = 1000$ (left) and $N_0 = 10000$ (right). For each network we compute the relative error for $\sigma = 0.1$, 1 and 10 by increments of 5% of growth. For small values of σ the error is lower, but in all settings, the reliability of the projection is high. Each point is calculated with 100 simulations.

where S^0 and S^z represent the matrices of weighted distances of the original network with N_0 nodes, and the grown network with size $N_z > N_0$, respectively.

Fig. 5 shows the local error measured on two growing networks by increments of 5% of growth. Their initial size is $N_0 = 1000$ (left) and $N_0 = 10000$ (right). For each network we compute the relative error for $\sigma = 0.1$, 1 and 10. When $\sigma = 0.1$ only the closest neighbors have a significant weight in the measurement of the local error. These low values of σ give the neighbor-wise error a very local sense. On the other hand, when $\sigma = 10$ the gaussian curve becomes flat and the measure is affected by the entire network perturbations, i.e. every node is equally considered as belonging to the neighborhood. Despite this global neighborhood for high σ values, the local error measure represents a relative distance to each node, and as we see, doubling the network size the average error remains below 0.1%. These very low error rates ensure a good reliability of the projection from a local point of view.

4. Conclusions

In this article we have raised the question about reliability of a standard linear projection technique such as SVD. The question is pertinent because SVD, and in particular its truncated version (TSVD), is rooted at the heart of some methodologies which pretend to extract useful and reliable information from dynamic data, i.e. data that is constantly undergoing change. We focus on growing scale-free networks.

We tackle the problem from two complementary points of view. At the large-scale level, we monitor average changes in nodes' TSVD projections. This means that each node's projection is compared against itself on successive changes.

Note however that success in practical applications of TSVD depends mostly on neighborhood stability. In other words, coherence of the output when data has suffered changes relies on the fact that the surroundings of a projected node are similar to those before those changes had happened. From a mathematical point of view, this merely implies that projections change in a coordinated way, such that relative positions are stable. Keeping this in mind, the local measure developed above captures this facet of the problem by comparing not the evolution of a nodes position against itself, but rather against the rest of nodes. Furthermore, we introduce a parameter to weight this variation depending on the distance from the node of interest. This tunable parameter allows for a finer observation of neighborhood stability, ranging from immediate neighborhood measures to far-reaching areas. Note that the local measure is orders of magnitude lower than the global one. This points to the fact that, although the projection changes significantly, displacements in the plane $\mathcal{U}_2$ are similar in magnitude and direction on average. In other words, as a node of the network grows following the preferential attachment, it is highly likely that its neighbors

also increase their weight staying close together.

Results indicate that TSVD projections are very robust against data growth. From a global point of view, an addition of 40% of new data implies only an average change of 10% from initial conditions. Doubling the amount of nodes to a network supposes a modification of 15% in the positions of the set of initial nodes. More importantly, changes at the local level (neighborhood) are close to 0 even in the most demanding case.

Such results have been obtained with rather large structures ($N_0 = 10^3$ and $N_0 = 10^4$), which at the end of the process have doubled their initial size. This ensures that TSVD is reliable in a wide range of situations. On the other hand, our study focuses on a particular network model (BA) in which time plays an important role: the later a node appears, the lowest its chances to become an important one (a hub). We anticipate that the irruption of important entities at late stages of evolution would surely disrupt TSVD projections in a more significant way. Nonetheless, we stress that growing systems typically develop smoothly, so our conclusions can be safely held.

Finally, we can briefly relate these results to the original motivation of the manuscript, that is, a scenario where the modular structure of networks is taken into account. In that situation, the stability of a TSDV map in the case of network changes is granted given the above reported results. Then, the characterization of the role of nodes and modules in terms of SVD's output can be safely regarded as faithful even in the case of severe changes in the underlying topology.

References

Arenas, A., Borge-Holthoefer, J., Gómez, S. & Zamora-López, G. [2010] "Optimal map of the modular structure of complex networks," *New Journal of Physics* **12**, 053009.

Barabási, A. & Albert, R. [1999] "Emergence of scaling in random networks," *Science* **286**, 509.

Berry, M., Dumais, S. & O'Brien, G. [1995] "Using linear algebra for intelligent information retrieval," *SIAM review* **37**, 573–595.

Capocci, A., Servedio, V., Colaiori, F., Buriol, L., Donato, D., Leonardi, S. & Caldarelli, G. [2006] "Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia," *Physical Review E* **74**, 36116.

Chu, M. & Golub, G. [2005] *Inverse eigenvalue problems: theory, algorithms, and applications* (Oxford University Press, USA), ISBN 0198566646.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. [1990] "Indexing by latent semantic analysis," *Journal of the American Society for Information Science* **41**, 391–407.

Golub, G. & Van Loan, C. [1996] *Matrix computations* (Johns Hopkins University Press, Baltimore, MD).

Landauer, T. & Dumais, S. [1997] "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge," *Psychol. Rev.* **104**, 211–240.

Landauer, T., Foltz, P. W. & Laham, D. [1998] "Introduction to Latent Semantic Analysis," *Discourse Process* **25**, 259–284.

Pastor-Satorras, R. & Vespignani, A. [2004] *Evolution and structure of the Internet: A statistical physics approach* (Cambridge Univ Pr), ISBN 0521826985.

Zlatić, V., Božičević, M., Štefančić, H. & Domazet, M. [2006] "Wikipedias: Collaborative web-based encyclopedias as complex networks," *Physical Review E* **74**, 16115.