Network Information Filtering for Big Data: a novel, fast, scalable and adaptive method

Tomaso Aste[1,2], Guido Previde Massara[1], Wolfram Barfuss[1,3] Rodrigo Mazorra[1] T. Di Matteo[4]

[1] Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK.
[2] Systemic Risk Centre, London School of Economics and Political Sciences, London, WC2A2AE, UK.
[3] Department of Physics, University of Erlangen-Nuremberg, DE.
[4] Department of Mathematics, King's College London, The Strand, London, WC2R 2LS, UK.

Abstract

We propose a new network-filtering algorithm - named TMFG - that uses any arbitrary similarity measure to gather complex, big dataset into a meaningful network structure that can be used for clustering, community detection, modeling and databasing. The method is scalable to very large datasets and it can take advantage of parallel and GPUs computing. The method is adaptable allowing online updating and learning with continuous insertion and deletion of new data as well as changes in the strength of the similarity measure. The approach consists in building a triangulation that maximizes a gain function associated with the amount of information retained by the network. We report applications to finance and big data analytics.

Keyword: Big Data, Network Filtering, PMFG, TMFG, Graphical modeling

1 Filtering information in big data by using networks

We are witnessing interesting times rich of information, readily available for us all. Using, understanding and filtering such information has become a major activity across science, industry and society at large. We need tools that can analyze this information while it is generated and providing ways to reduce complexity and dimensionality while keeping the integrity of the dataset. Information content and flow are often associated with large degrees of redundancy. Redundancy is often used to convey strength to the meaning or, more simply, it is the signal of recurring patterns with high statistical significance and therefore important. In this presentation we propose to use such redundancy to build an information-based network that retains the relevant part of the data-interdependency structure. The structure of this network is a representation of the information in the dataset and such information can be efficiently analyzed by using network-theoretic tools.

The idea of using redundancy - namely correlation coefficients - to filter information in large-scale datasets by building networks of relevant links has been very actively studied in the literature mostly by means of two approaches: 1) the minimum spanning tree (MST) [1, 2] and 2) the planar maximally filtered graph (PMFG) [3, 4]. The common idea underneath these two approaches is to retain the largest and most significant possible sub-graph while imposing global constraints on the topology of the resulting network. In particular, in the MST approach, the links with largest weights are retained while constraining the sub-graph to be globally a (spanning) tree. Similarly, in the PMFG construction the largest weights are retained while constraining the sub-graph to be globally a planar graph. The PMFG has richer information content than the MST with a larger number of edges (3N-6 instead of N-1, with N being the number of vertices) and the presence of 3- and 4-cliques.

2 Planar Information Filtering Graphs

PMFGs are powerful tools to study complex datasets. For instance, it has been shown in [5] that by making use of the 3-clique structure of the PMFG a clustering can be extracted allowing dimensionality reduction that keeps both local information and global hierarchy in a deterministic manner without the use of any prior information. Applications to financial data-sets can meaningfully identify industrial activities and structural market changes [6] and can be used to diversify financial risk by building a well-diversified portfolio that effectively reduces investment risk. Specifically investments in stocks that occupy peripheral, poorly connected regions in the financial filtered networks are most successful in diversifying invest-

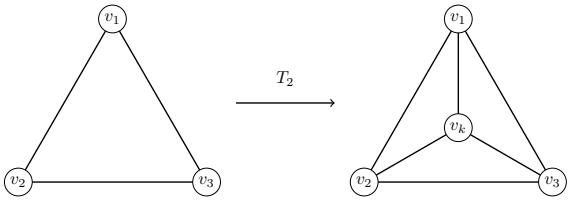

Figure1: T_2 move: addition of one vertex inside a triangular face [8, 9].

ments even for small baskets of stocks [7].

However, the algorithm so far proposed to construct the PMFG is numerically costly with $O(N^3)$ computational complexity and cannot be applied to large-scale data. Here we introduce a new algorithm, the TMFG (Triangulated Maximally Filtered Graph), that efficiently extracts a planar subgraph which optimizes some objective function (which we shall call "gain function"). The TMFG algorithm [10], outlaid below, starts from a triangle and adds vertices inside triangles (local move T_2 [8, 9] see Fig.1). The novelty is that, at each step, the algorithm optimizes a *gain function* $G(v_k, \{v_a, v_b, v_c\})$ that quantifies the gain achievable by adding vertex v_k inside the triangle $\{v_a, v_b, v_c\}$.

```
   input : W — a similarity matrix
   output: TMFG — a filtered version of W respecting the planarity constraint
   /* Initialise a triangle t_1 e.g. by using the highest W                */
 1 t_1 ← Three vertices with highest W ;
 2 VertexList ← List of vertices of W not belonging to t_1 ;
 3 Calculate Gains(VertexList, t_1) ;
 4 N ← number of vertices in VertexList ;
   /* Insert N − 3 vertices via T_2                                        */
 5 for n = 1 to N − 3 do
 6 |   (v_i, t_abc) = argmax_{v_k, t_xyz} {Gains(v_k, t_xyz)} ;
 7 |   eliminate row Gains(v_i, :) ;
 8 |   eliminate column Gains(:, t_abc) ;
 9 |   t_a1, t_a2, t_a3 ← triangles created by the insertion of v_i ;
10 |   update gain matrix Gains(:, t_a1), Gains(:, t_a2), Gains(:, t_a3) ;
   |   /* Execute T_1                                                      */
11 |   Evaluate Gain by implementing T_1 over t_a1, t_a2, t_a3 and their neighbors and execute
   |   T_1 if net gain is positive ;
12 |   if T_1 is executed then
13 |   |   Evaluate Gain by implementing T_1 over modified triangles and their neighbors ;
14 |   end
15 end
16 return TMFG ;
```

TMFG algorithm pseudocode [10].

The TMFG graph is very similar to the PMFG and it can be successfully applied to the same domains. For instance, an example of the use of the TMFG to extract the industrial sectors structure from correlations is reported in Fig.2, this is consistent with the results in [6] obtained with the PMFG. The advantage is that the TMFG algorithm is much more efficient computationally. It is scalable to very large datasets , given its local nature, it is ideally suited for parallelisation. The algorithm has the advantage of allowing 'online' updates of the planar graphs through simple local moves. It can be naturally applied to multipoint dependency measures taking advantage of

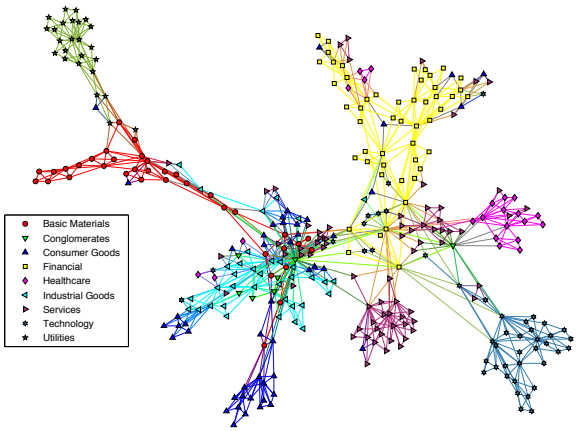

Figure2: Example of TMFG built from cross-correlations among daily log-returns of 342 US stocks across a period of 15 years (1997-2012) [6].

the 3- and 4-clique structure. Further, it is not restricted to planar topologies allowing higher-genus hyperbolic embeddings to be explored [3, 9]. Finally, another appealing advantage concerns graphical modeling (e.g. Markov Random Fields) where the structure of the network ensures that exact inference algorithms can be performed in an efficient fashion.

References

[1] R. C. Prim, Bell-System Technical Journal 36 (1957) 1389-1401.

[2] R. N. Mantegna., Eur. Phys. J. B 11 (1999) 193-197.

[3] T. Aste, T. Di Matteo, S. T. Hyde, Physica A 346 (2005) 20.

[4] M. Tumminello, T. Aste, T. Di Matteo, R. N. Mantegna, PNAS 102, n. 30 (2005) 10421.

[5] W. M. Song, T. Di Matteo, and T. Aste, PLoS ONE 7 (2012) e31929.

[6] N. Musmeci, T. Aste, T. Di Matteo, arXiv:1406.0496 [q-fin.ST] submitted 2014.

[7] F. Pozzi, T. Di Matteo and T. Aste, Scientific Reports 3 (2013) 1665.

[8] T. Aste and D. Sherrington, J. Phys. A: Math. Gen. 32 (1999) 7049-56.

[9] T. Aste, R. Gramatica, and T. Di Matteo Phys. Rev. E 86 (2012) 036109.

[10] G. Previde Massara, T. Di Matteo and T. Aste, draft (2014).