Conceptual and Statistical Footprints for Social Networks’ Characterization

Trad Mohamed Riadh
Laboratoire MAS, Ecole Centrale Paris
tradriadh@hotmail.com

Bénédicte Le Grand
Laboratoire d’Informatique de Paris 6 - UPMC
Benedicte.Le-Grand@lip6.fr

Marie-Aude Aufaure
Laboratoire MAS, Ecole Centrale Paris
Marie-Aude.Aufaure@ecp.fr

Michel Soto
Laboratoire d’Informatique de Paris 6 - UPMC
Michel.Soto@lip6.fr

ABSTRACT
This article proposes a method relying on Formal Concept Analysis and Galois lattices for complex systems analysis. Statistics based on concept lattices enable the computation of the “Conceptual Distribution” of objects classified by the lattice. Experimentation on sample datasets extracted from three online social networks illustrates the use of these conceptual statistics for the global characterization and the automatic filtering of these systems. Moreover, compared to classical measures, these statistics offer new perspectives for object filtering and lattices simplification that would be useful for lattices visualization and interpretation. However, conceptual statistics calculation, based on Galois lattices computation requires expensive calculations. This paper focuses on conceptual statistics contribution and optimized methods for their calculation for scalability purposes.

1. INTRODUCTION

This paper proposes a method relying on Formal Concept Analysis and Galois lattices for Complex Systems analysis. This technique provides an intuitive visual characterization of the systems under study through the computation of Conceptual Footprints generated from Galois lattices. These footprints (defined in Section 3.2) help users better understand the data’s overall structure and features, as well as identify significant elements. This method also enables the automation of outliers’ filtering.

Although this approach is generic and may therefore be applied to any type of complex system, it is illustrated in this paper in the context of social networks. Online social networking sites such as Myspace, Facebook, Flickr or LinkedIn have recently achieved a tremendous success. With such sites, social networks can be built, based on professional relationships, common studies, specific hobbies, etc. These social systems have become the most popular way of sharing content, express opinions and create relationships with other users. Social search, social navigation as well as social visualization have become challenging tasks in this context. Social network analysis[12] is related to the understanding and interpretation of the network behavior. Analysis also provides useful information about the way communities are formed and interact. Social networks have been studied from mathematical and statistical points of view[9], and also in computer science to provide methods for social search, navigation and visualisation[11]. An interesting way to understand and interpret social networks interaction is to combine analysis and visualization as this is done in[13] with Pajek software. The techniques proposed in this paper constitute an alternative approach to centrality measures for the analysis of these networks.

The major contribution of this work is to introduce new methods for complex system analysis based on Formal Concept Analysis and Galois lattices and their contribution over existing methods such as centrality-based methods. Although Formal Concept Analysis, based on Galois lattices adds semantics upon analyzed datasets, its computational cost is still high and proportional to the size of studied datasets. One can think about reducing the initial dataset size, removing objects that would not affect the analysis. However, classical measures for network analysis focus on the graph structure regardless of any semantics, hence the need for new measures for identifying these objects.

This article is organized as follows. Section 2 describes

1http://www.myspace.com
2http://www.facebook.com
3http://www.flickr.com
4http://www.linkedin.com
centrality measures used in Complex Systems analysis. In Section 3, we introduce our conceptual measures. Section 4 presents results for the conceptual analysis over real datasets as well as an evaluation of the contribution of this work with regard to traditional centrality measures. Finally, Section 5 explores a distributed approach for Conceptual Analysis.

2. CENTRALITY MEASURES

In graph theory, there are several measures that determine the relative importance of a vertex within a graph. However, quantifying the importance of a vertex depends on the analysis purpose, thus there are diverse definitions for centrality.

In data analysis, there are standard measures of centrality including closeness[8], betweenness[7], degree and stress centrality.

In the experimentation described in section 4, we will focus on the first two measures, let C_b and C_c denote respectively the betweenness and closeness centrality.

Let G = <V,E> be a finite graph and let V and E be the set of vertices and edges of this graph.

We use \( d_G(s,t) \) to denote the length of the shortest path between vertices s and t. By definition, \( d_G(s,s) = 0 \) for every \( s \in V \), and \( d_G(s,t) = d_G(t,s) \) for \( s,t \in V \).

Let \( \sigma_{st} = \sigma_{ts} \) denote the number of shortest paths from \( s \in V \) to \( t \in V \), where \( \sigma_{ss} = 1 \) by convention. Let \( \sigma_{sv} \) denote the number of shortest paths from \( s \) to \( t \) containing \( v \) where \( v \in V \). Closeness and betweenness centrality are expressed as follows.

- **Closeness centrality**

  \[
  C_c(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}
  \]

- **Betweenness centrality**

  \[
  C_b(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{sv}}{\sigma_{st}}
  \]

High closeness centrality indicates that a vertex can reach others on relatively short paths whereas high betweenness centrality indicates that a vertex lies on considerable fractions of shortest paths connecting others.

One can also focus on relations between these vertices, thus centrality measures can be extended for quantifying the importance of an edge within a graph.

However, these measures do not take into account the semantic relationship between the different objects, they only focuses on the graph structure. In the following section we will introduce new measures based on Galois lattices and Formal Concept Analysis.

3. SOCIAL NETWORKS’ CONCEPTUAL ANALYSIS

3.1 Formal Concept Analysis and Galois Lattices

FCA is a mathematical approach to data analysis which provides information with structure. FCA may be used for conceptual clustering as shown in[5] and[14]. The notion of Galois lattice to describe a relationship between two sets is the basis of a set of conceptual classification methods. This notion was introduced by[3] and[2]. Galois lattices group objects into classes according to the properties they have in common.

Consider two finite sets \( D \) (a set of objects) and \( M \) (the set of these objects’ properties also called attributes), and a binary relation \( R \subseteq D \times M \) between these two sets. Let \( o \) be an object of \( D \) and \( p \) a property of \( M \). We have \( oRp \) if the object \( o \) has the property \( p \).

Let \( P(D) \) be the powerset of \( D \) and \( P(M) \) the powerset of \( M \) Each element of the lattice is a couple, also called concept, noted \( (O,A) \). A concept is composed of two sets \( O \in P(D) \) and \( A \in P(M) \) which satisfy the two following properties (1):

\[
\begin{align*}
A & = f(O) \text{where } f(O) = \{ a \in M | o \in O, oRa \} \\
O & = f'(A) \text{where } f'(A) = \{ o \in D | a \in A, oRa \}
\end{align*}
\]

\( O \) is called the concept’s extent, and \( A \) is its intent. The extent is a subset of the system’s objects and the intent is a subset of the system’s properties.

Galois lattices’ complexity in size, due to the very high number of concepts they contain, makes them very difficult to interpret with traditional Hasse diagrams. In [10], the authors have have defined interest measures to reduce the size of large concept lattices and apply their method to health care social communities. The approach proposed in this paper consists in computing statistics which lead to the definition of a Conceptual Distribution, presented in the following Section.

3.2 Social Networks’ Conceptual Distributions

The Galois lattice is used to calculate statistics for every object of the social network. Within a social network, objects are members and their properties may be their contacts. In the following, let an object \( o \).

- **Relatedness**

  Let \( C \) the set of concepts from the lattice which contain the object \( o \) in their extents. Let \( C' \) the subset of concepts from \( C \) which contain at least another object than \( o \) in their extents and at least one property in their intents. The value of \( o \)’s Relatedness (noted Relatedness(\( o \))) corresponds to the average number of objects with which \( o \) is clustered in concepts from \( C' \), divided by the total number of objects in the social network. The Relatedness value indicates whether \( o \) is connected to many other objects.

- **Closeness**

  Let \( S \) the set of objects with which \( o \) is clustered in at least one concept of the lattice (i.e. the set of objects
with which \(o\) is connected); these objects have at least one property in common with \(o\) (by construction). The object \(o\)'s \textit{Closeness} value (noted \textit{Closeness}(\(o\))) is the average number of properties \(o\) shares with the other objects from \(S\), divided by the total number of \(o\)'s properties. This parameter indicates whether \(o\) is strongly connected to the other objects from \(S\) (i.e. whether it shares a high proportion of properties with them).

- **Conceptual Distribution**
  The \((\textit{Relatedness}(o), \textit{Closeness}(o))\) pair constitutes the **Conceptual Distribution** of the object \(o\).

- **Conceptual Footprint**

\[
\text{conc}\_\text{footprint}(syst) = \frac{\text{relatedness}(syst)}{\text{closeness}(syst)}
\]

\[
\text{relatedness}(syst) = \frac{\sum_{i=1}^{N} \text{relatedness}(o_i)}{N}
\]

\[
\text{closeness}(syst) = \frac{\sum_{i=1}^{N} \text{closeness}(o_i)}{N}
\]

The average values of the \textit{Relatedness} and \textit{Closeness} parameters for all objects of a given system represent the Conceptual Footprint of this system. Let \(syst\) the complex system under study, comprising \(N\) objects.

4. EXPERIMENTATION AND RESULTS

4.1 **Description of the Sample Datasets**

For this experimentation, four social networks samples have been collected with a dedicated crawler. This crawler parses networks from given starting points (i.e. members of the social network), retrieves the contacts of these “initial” members and looks for the contacts of their contacts recursively. This contacts’ extraction is bounded by two parameters defining respectively the depth of the search and the maximum number of contacts retrieved for each member. Each crawled member, i.e. for whom the contacts have been retrieved, is called an object, and his/her contacts correspond to his/her properties. In the collected samples, the total number of members is much higher than the number of objects as only a portion of them have been parsed (due to depth’s restriction). The four samples’ features are summarized in Table 1.

These samples are partial and biased views of the studied social networks, as they only contain subsets of members, selected with various criteria in terms of starting points for the crawling process, depth and maximum number of contacts. The interpretations provided in this paper are only valid locally, for the “regions” from which the datasets are extracted from. As shown in the following, different results may be observed with the two samples collected from the Flickr network.

4.2 **Conceptual Analysis’ Results**

The results presented in this section illustrate how the conceptual approach may be used for social networks’ global characterization as well as outliers’ filtering.
values’ homogeneity. Flickr50’s Relatedness and Closeness values are indeed much more homogeneous (as they are gathered in the same region of the graphic) than Myspace’s, where several clusters may be distinguished.

4.2.2 Social Networks’ Outliers Filtering

The Conceptual Distributions of Flickr5 and DailyMotion show members Relatedness and Closeness values are very low (See Figure 2). These objects are called marginal and it is interesting to compute the new Conceptual Distributions obtained after eliminating these outliers (i.e. after deleting the bottom left bubble).

The results are presented on Figure 3 and Figure 4. After this filtering process, DailyMotion’s Conceptual Distribution (Figure 3) no longer contains any outliers “bubble”, whereas Flickr5’s ‘Conceptual Distribution’ (Figure 4) has new marginal elements. If this filtering process is carried on with the Flickr5 dataset by computing the Conceptual Distribution again after the elimination of this new bubble, other marginal elements are still present. The Flickr5 sample is therefore intrinsically heterogeneous; it always contains outliers, whatever the number of filtering steps.

The filtering process has been automated by eliminating from the original dataset all objects whose average Relatedness and Closeness values are below the average values for the overall network minus $\alpha \times$ standard deviation for each parameter, where $\alpha$ may vary.

Relatedness and Closeness parameters are computed again for all objects of the new sample then outliers are eliminated, and so on. The automatic filtering algorithm converges when the dataset no longer contains any marginal object, i.e. when the Relatedness and Closeness values of all remaining objects are sufficiently homogeneous (or when there is only one object left if the system is highly heterogeneous).

Figure 5 compares this conceptual filtering applied to the four studied networks with $\alpha = 0.25$. As expected, the filtering process based on conceptual parameters strongly affects Flickr5 network as some marginal elements always remain after the various filtering steps. On the other hand, no member from Myspace network is eliminated.

4.3 Centrality-based filtering

As for Conceptual Filtering, Centrality-based Filtering consists in eliminating from the original dataset all objects whose average betweenness and closeness centrality values are below the average values for the overall network minus $\alpha \times$
standard deviation for each parameter, where $\alpha$ may vary.

There are different algorithms for centrality computation. Most of them have a $\theta(n^3)$ time complexity, where $n$ is the number of vertices in the network. In [4] Brandes introduces a more efficient algorithm based on a new accumulation technique that reduces complexity to $\theta(nm)$ time complexity where $m$ is the number of edges which we use in our experimentation for centrality measures computation.

Since computing the Conceptual Distribution for an object $o$ requires the calculation of the whole lattice, centrality-based filtering seems to be faster than the conceptual one. In the following section, the two approaches are compared in order to understand their respective impacts on resulting lattices.

### 4.4 Comparison between Centrality-based and Conceptual filtering

In order to achieve this comparison, Conceptual and Centrality-based filtering are applied on the same dataset (Flickr50). The original Conceptual Distribution of this sample is represented on Figure 6.

The Galois lattices associated to the new set of objects (after the Conceptual and Centrality-based filtering) are computed for different values of $\alpha$.

Figures 7, 8 and 9 represents the Conceptual Distributions obtained after the conceptual filtering for $\alpha = 1.25$, $0.6$ and $0$ respectively whereas figures 10, 11 and 12 represents the Conceptual Distributions obtained after Centrality-based filtering for the same $\alpha$ values. In each graph, we represent the original distribution (in red) and the distribution after the filtering process (blue and yellow respectively for Conceptual Filtering and Centrality-based Filtering) in order to evaluate the impact of the filtering process.

For high $\alpha$ values, all nodes are kept, which is not surprising as high $\alpha$ values imply very low thresholds. Therefore, Conceptual Distributions for both filtering process remain the same as illustrated in figures 7 and 10.

As $\alpha$ decreases, some nodes are filtered. Figures 8 and 11 show the new Conceptual Distributions after the conceptual respectively the centrality-based filtering.

Conceptual filtering eliminates vertices according to their relevance into the Galois lattice thereby discarding vertices which are not strongly connected to the other vertices (closeness) and outliers (relatedness) from a conceptual point of view.
Centrality-based filtering eliminates vertices according to their centrality measures which generally determine the relative importance of a vertex within the graph regardless of any semantics.

As expected, Conceptual filtering preserves the most significant “conceptual” vertices, i.e. objects which have the highest impact on the Conceptual Distribution of the remaining objects. This is not the case with Centrality-based filtering which may discard objects with high conceptual parameters values, leading to residual “bottom-left corner bubbles” in Conceptual Distributions.

Figures 9 and 12 confirm this result for lower $\alpha$ values.

5. SCALABLE AND DISTRIBUTED APPROACH FOR CLOSENESS AND RELATEDNESS COMPUTING

Although Conceptual Filtering can bring real benefits to social network analysis as shown in this paper, its computational cost is high. Indeed, conceptual filtering is based on closeness and relatedness computation implying the whole lattice calculation. Therefore, in this section we introduce a new approach for closeness and relatedness computation based on the ELL Algorithm[6].

Let $o$ be an object of $D$ and let $C$ the set of concepts from the lattice which contain the object $o$ in their extents. The calculation of the Conceptual Distribution associated to $o$ is based on the determination of $C$. 

Figure 6: Original distribution for Flickr50 dataset

Figure 8: Conceptual Filtering $\alpha = 0.60$

Figure 9: Conceptual Filtering $\alpha = 0.00$
Our approach for optimizing Conceptual Distribution computation is based on the computation of the subset C instead of the whole lattice which is possible using the ELL Algorithm.

For two disjoint subsets \(X_0\) and K of the set of objects I, the ELL Algorithm lists all the closed sets of I obtained by extending \(X_0\) with some elements of K i.e. all the closed sets which strictly contain \(X_0\) and are contained in \(X_0 \cup K\).

Moreover, Baklouti et al. introduced a distributed and scalable version of the ELL Algorithm in [1] for C computation called SD-ELL. Although this approach would substantially optimize the Conceptual Distribution computation, it requires suitable infrastructure for job distribution.

One might think about allocating a set of objects for each node of the infrastructure i.e. each computational resource within a grid. Different approaches are being studied and will be subject for future publications.

6. CONCLUSION
This article presented a conceptual method for Galois lattices analysis and interpretation. This conceptual characterization relies on parameters, called Relatedness and Closeness, computed for each object according to the extents and intents of the lattices’ concepts. The Conceptual Distributions and footprint deduced from these values enable the systems’ global characterization in terms of homogeneity or heterogeneity, as well as an automatic filtering of outliers, called marginal objects. The proposed generic method has been applied to four social networks samples in order to illustrate its operation and results. Moreover, the filtering process, based on this conceptual characterization, proves to be more efficient in preserving objects that would affect the lattice structure than filtering methods based on traditional centrality measures. Deleting these nodes would affect slightly the lattice structure and the resulting lattice would be more similar to the original one than lattices resulting from centrality-based filtering. However, computing Conceptual Distribution for all objects within the graph is very costly in terms of time and space complexity. Therefore, perspectives of this work consist in investigating different approaches for Conceptual Distribution computing optimization.

7. REFERENCES


