On the Phenomenological Reconstruction of Complex Systems—The Scale-Free Conceptualization Hypothesis

Gonzalo A. Aranda-Corral[1], Joaquín Borrego-Díaz[2] and Juan Galán-Páez[2*]

[1] *Department of Information Technology, Universidad de Huelva, Palos de La Frontera, Spain*
[2] *Department of Computer Science and AI, Universidad de Sevilla, Sevilla, Spain*

Phenomenological reconstruction of a complex system (CS) from collected and selected data allows us to work with formal models (representations) of the system. The task of building a qualitative model necessitates the formalization of relationships among observations and con-crete features. Formal concept analysis can help to understand the conceptual structure behind these qualitative representations by means of the so-called *concept lattices* (CLs). The study of these kinds of semantic networks suggests that a strong relationship exists between its topological structure and its soundness/usefulness as a qualitative representation of the CS. The present paper is devoted to this question by presenting the so-called *scale-free conceptualization hypothesis*. The hypothesis claims that a scale-free distribution of node connec-tivity appears on the CL associated to complex systems (CLCS) only when two requirements hold: CLCS is useful both to represent qualitative and reliable attributes on the CS, as well as to provide a basis for (qualitatively) successfully reasoning about the CS. Experiments revealed that the topologies of CLCS are similar when the amount of information on the CS is sufficient, whereas it is different in other CLs associated to random formal contexts or to other systems in which some of the former requirements do not hold.

Keywords complex systems; complex networks; formal concept analysis; scale free topology; scale free conceptualization hypothesis

INTRODUCTION

Complex system (CS; and complex network) is a broad concept, which has specific features but covers very different systems, with an astonishing variety of dynamics. Among them, particularly interesting are those related to human (rational) activities, such as organizations, communities and cities. An interesting feature of CS research is investigating how humans describe and understand such types of systems. It is very intriguing how the rationality of humans is able to select important

*Correspondence to: Juan Galán, Department of Computer Science and AI, Universidad de Sevilla, Sevilla, Sevilla, Spain.
E-mail: juangalan@us.es

features and concepts (and the relationships among them) of the CS in order to reason and predict its dynamics (for surviving within) or to describe their features.

This paper presents an empirical analysis of a kind of semantic networks built to specify the (qualitative) knowledge retrieved from CS. Semantic networks represent a useful tool to represent specific features of CS, such as language evolution and structure (e.g. Motter *et al.*, 2002), but their scope may include other CS where the analysis of semantic relationships among concepts involved in CS is necessary. General (formal) representations of semantic relationships are often complex networks as well, and their analysis provides some insights on the complex nature of the CS whose features are represented.

In the representation of CS, it is necessary to choose what features the observer thinks are essential to understand it, as well as how these features are related. Human reasoning activities (Simon, 1982) are essential to understand the behavior of a system, and they strongly depend on our ability to quickly select key features and reason with limited resources. This reasoning cannot use every existent relationship among features; only the few the user believes are important (a concrete *bounded rationality* strategy, e.g. Goldstein and Gigerenzer, 2002). In fact, humans have bounded reasoning (BR) methods to carry out two interesting tasks in which CS research is particularly interested (Bourgine *et al.*, 2009):

- Identifying relevant entities for a given time and space scale.
- Characterizing interactions between entities.

By using knowledge representation and reasoning methods, the results of these tasks are useful for assessing and formalizing the system behavior. As was already mentioned in (Bourgine *et al.*, 2009), such Computer Science methods are essential to provide exploratory tools for a data-based approach to the problem.

In contrast to the former BR skills, global exploratory tools are faced with different problems. Formal Epistemology succinctly provides a general framework where explanation, simulation and validation comprise a number of tools and formalisms (see Figure 1 extracted from Bourgine *et al.*, 2009). In Figure 1, three sets of activities have been highlighted: the first region is devoted to the theoretical reconstruction (qualitative, in our case) of the system, and the second region is devoted to the simulation process of the system, which allows the reconstruction of the dynamics (in our case, by means of qualitative reasoning). In the last region (experimental validation), simulation and reconstruction are evaluated.

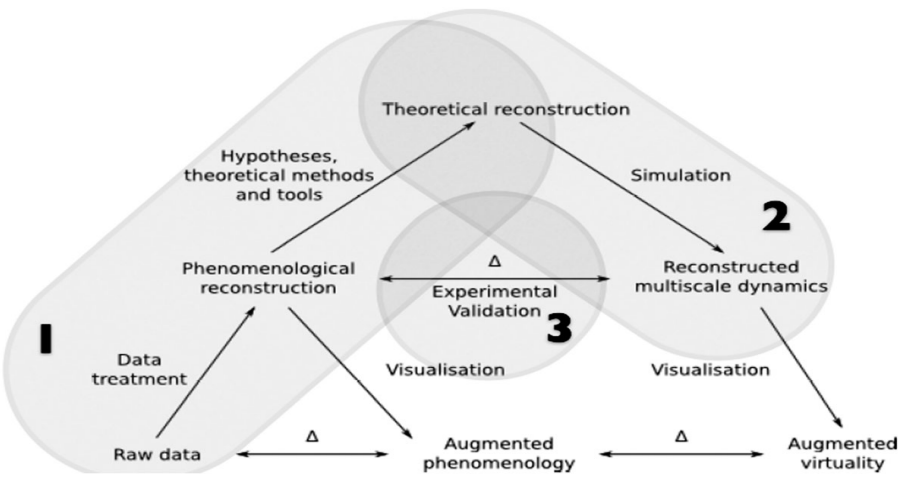

Figure 1 Formal and applied epistemology activities involved in the simulation of bounded rationality human tasks on complex system (from Bourgine et al., n.d.*)*

In summary, exploiting BR techniques in phenomenological reconstruction strongly depends on a sound selection of the features (qualitative) to be computed/studied. Therefore, some questions arise:

- How is it decided if the feature selection is sound?
- How is the soundness of the qualitative representation/reasoning model obtained from a feature selection analysed?
- Specifically: do sound qualitative modelizations share similar structure/properties?

The last question is particularly interesting, because a positive answer could aid in solving the two first ones.

Independently from the use of feature selection to obtain reliable systems, it is interesting to analyse the full information available. The global structure, which comprise of all the features, concepts and their relationships, represents a complex structure, which can provide more information about the CS than user deductions. Examples where interest in the global complex structures require more attention are stock markets, economy, ontology/folksnomy evaluation, betting markets and, in general, those CS in which humans are involved and where their decisions are based on partial information about the CS. To model this semantic network, formal concept analysis (FCA) is selected as a knowledge representation and reasoning framework.

The Role of Formal Concept Analysis in the Approach

Formal concept analysis (Ganter and Wille, 1999) is a mathematical theory for data analysis whose basic data structure is the formal context, consisting of a set of objects and their properties. They represent weak structures easily built from experience that allows the extraction of knowledge. Despite their simple data structure, formal contexts are useful structures for knowledge extraction and reasoning (cf. Ganter and Wille, 1999). When FCA is applied on a considerable amount of observable features of the CS, the concept lattice (CL) generated includes every concept involved in CS description; it represents a complex (semantic) network. Concepts and relationships within the CL can describe features of the CS as well as aid the observer tin working with the information (for describing, classifying, predicting, etc.). Authors have used FCA to study specific CS (Aranda-Corral *et al*., 2013) as well as to design automated processes of knowledge conciliation (Aranda-Corral and Borrego-Díaz, 2010; Aranda-Corral *et al*., 2012). FCA is used in this paper to support activities involved in the Regions 1 and 2 of Figure 1.

Aim of the Paper

The aim of this paper is twofold. On one hand, we discuss the main issues related to CLs associated with observations on CS. We will see how the analysis of topological features of CL associated to CS (CLCS), considered as complex networks, may in some cases aid understanding CS evolution. On the other hand, we assert that when the observations are objective and relevant in order to study the CS, the associated CL exhibits a scale-free distribution structure. This claim (which we call *scale-free conceptualization hypothesis*, SFCH) is experimentally analysed in a number of examples that cover several CS where different types of BR are used (both in the study and by agents within the CS).

Structure of the Paper

The structure of the paper is as follows. The next section is devoted to succinctly presenting FCA. Sections on Formal Concept Analysis Based Reasoning on Complex System and Scale-Free Residue of Concept Lattices describe how to associate CLs to CS from observations and assert SFCH. Other sections are devoted to the analysis of different activities associated to qualitative analysis of CS: forecasting/prediction (Forecasting/Prediction section) and catalogation/classification of objects (Classification section). Examples of Semantic systems (subsystems of WordNet) are analysed in WordNet Subsystems section. In Random Contexts section, the contrast with the topology associated to CLs extracted

from random contexts is explicitly presented. Lastly, the final section provides conclusions and describes future work (Concluding Remarks and Future Work section).

BACKGROUND: FORMAL CONCEPT ANALYSIS

Formal Concept Analysis mathematizes the philosophical understanding of a concept as a unit of thoughts composed of two parts: the extent and the intent. The extent covers all objects belonging to the concept, whereas the intent comprises all common attributes valid for all the objects under consideration (Ganter and Wille, 1999). It also allows the computation of concept hierarchies from data tables.

A *formal context* $M=(O,A,I)$ consists of two sets, O (objects) and A (attributes), and a relation $I \subseteq O \times A$. Finite contexts can be represented by a 1-0 table (identifying I with a Boolean function on $O \times A$). Given $X \subseteq O$ and $Y \subseteq A$, it defines

$$X' = \{a \in A | oIa \text{ for all } o \in X\}$$
$$Y' = \{o \in O | oIa \text{ for all } a \in Y\}$$

The main goal of FCA is the computation of the CL associated with the context. A (formal) concept is a pair (X,Y) such that $X'=Y$ and $Y'=X$. For example, the CL from the formal context of fishes of Figure 2, left (attributes are understood as 'live in') is depicted in Figure 2, right. Each node is a concept, and its intension (or extension) can be formed by the set of attributes (or objects) included along the path to the top (or bottom). For example, the bottom concept ({*eel*}, {*Coast, Sea, River*}) is the concept *euryhaline fish*. CL contains every concept that can be extracted from the context. As well, concepts are defined, but it is possible that no specific term (word) exists to denote it.

Logics for Formal Concept Analysis: Implications and Basis

Logical expressions in FCA are *implications between attributes*. An implication is a pair of sets of attributes, written as $Y_1 \to Y_2$, which is true with respect to $M=(O,A,I)$ according to the following definition.

A subset $T \subseteq A$ *respects* $Y_1 \to Y_2$ if $Y_1 \not\subseteq T$ or $Y_2 \subseteq T$. It says that $Y_1 \to Y_2$ holds in M ($M \models Y_1 \to Y_2$) if for all $o \in O$, the set $\{o\}'$ respects $Y_1 \to Y_2$. In that case, it is said that $Y_1 \to Y_2$ is *an implication* of M.

Definition 2.1: Let $\mathcal{L}$ be a set of implications and L be an implication.

(1) L follows from $\mathcal{L}$ ($\mathcal{L} \models L$) if each subset of A respecting $\mathcal{L}$ also respects L.
(2) $\mathcal{L}$ is complete if every implication of the context follows from $\mathcal{L}$.
(3) $\mathcal{L}$ is non-redundant if for each $L \in \mathcal{L}$, $\mathcal{L} \setminus \{L\} \not\models L$.
(4) $\mathcal{L}$ is (an implication) basis for M if $\mathcal{L}$ is complete and non-redundant.

	River	Coast	Sea
Carp	X		
Escatofagus	X	X	
Bream		X	X
Sparus		X	X
eel	X	X	X

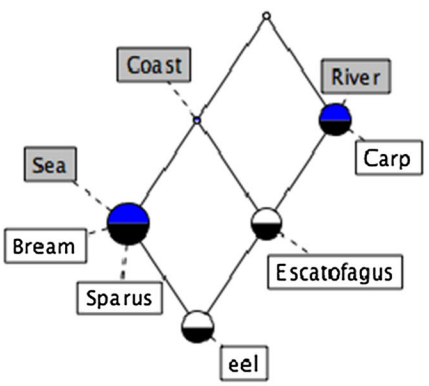

Figure 2 Formal context of fish and its concept lattice

It can obtain a basis from the *pseudo-intents* (Guigues and Duquenne, 1986) called *stem basis* (SB). An SB for the formal context of living beings is provided in Figure 2 (down). It is important to remark that SB is only an example of a basis for a formal context. In this paper, no specific property of the SB is used, so it can be replaced by any implication basis.

In order to work with formal contexts, SB and association rules, the CONEXP[1] software has been selected. It is used as a library to build the module that provides the implications (and association rules) for the reasoning module of our system. The reasoning module is a production system (designed for Aranda-Corral and Borrego-Díaz, 2010). Initially it works with SB, and entailment is based on the following result:

Theorem 2.2: Let $\mathcal{S}$ be a basis for M and $\{A_1, \ldots, A_n\} \cup Y \subseteq A$. The following conditions are equivalent:

1) $\mathcal{S} \cup \{A_1, \ldots A_n\} \vdash_p Y$ ($\vdash_p$ is the entailment with the production system).
2) $M \models \{A_1, \ldots A_n\} \rightarrow Y$.

Stem basis is an adequate knowledge base (KB) for the production system to reason about attributes and concepts. However, SB is designed to entail true implications only, without any exceptions to the object set nor implications with a low number of counterexamples in the context.

Another, more important question arises when working on predictions. In this case, the goal is to obtain methods for selecting a result among all entailed conclusions (eventually they are mutually incoherent), and Theorem does not provide such a method. Therefore, it is better to consider association rules (with confidence) instead of true implications, and the initial production system must be revised for working with confidence.

Researching logical reasoning methods for association rules is a relatively recent and promising line of research (Balcázar, 2010). In FCA, association rules are implications between sets of attributes. Confidence and support are defined as usual. Recall that the *support* of X, $supp(X)$ of a set of attributes X is defined as the proportion of objects that satisfy every attribute of X, and the *confidence* of an association rule is $conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$. Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of an object satisfying every attribute of Y under the condition that it also satisfies every one of X. CONEXP software provides association rules (and their confidence) for formal contexts (called *Luxenburger basis* Luxenburger, 1991).

FORMAL CONCEPT ANALYSIS BASED REASONING ON COMPLEX SYSTEM

Complex networks are a widely used representation of selected features from a CS. The topological structure of the network aids in understanding a considerable number of characteristics of the associated CS. When the goal is to reason with qualitative features, it can be interesting to extract emergent concepts from these interactions. It is here where FCA can play a relevant role.

The selection of FCA for processing qualitative information about CS lies in the fact that human reasoning—in fact, our BR skills—about the dynamics and organization of a CS has a qualitative nature. Therefore, human reasoning and conjectures about the CS can be expressed in qualitative terms (possibly choosing thresholds and multivalued attributes). Once qualitative hypothesis are presented, even non-symbolic mechanisms for reasoning can be useful to validate the conjectures. The qualitative reasoning process by means of FCA would be depicted as in Figure 3. The observer has to select attributes and objects they consider relevant to determine CS dynamics, and the reasoning focuses on the associated subcontext by selecting interesting attributes (contextual selection). Then he/she can consider the elements of CS as objects of a formal context. This context (often with a huge size) is built by means of data extraction and processing, expert observations, data mining and so on. It is expected that reasoning with the contextual selection provides some information about the CS. If the goal is to reason with qualitative features, it is interesting to extract

[1] http://sourceforge.net/projects/conexp/

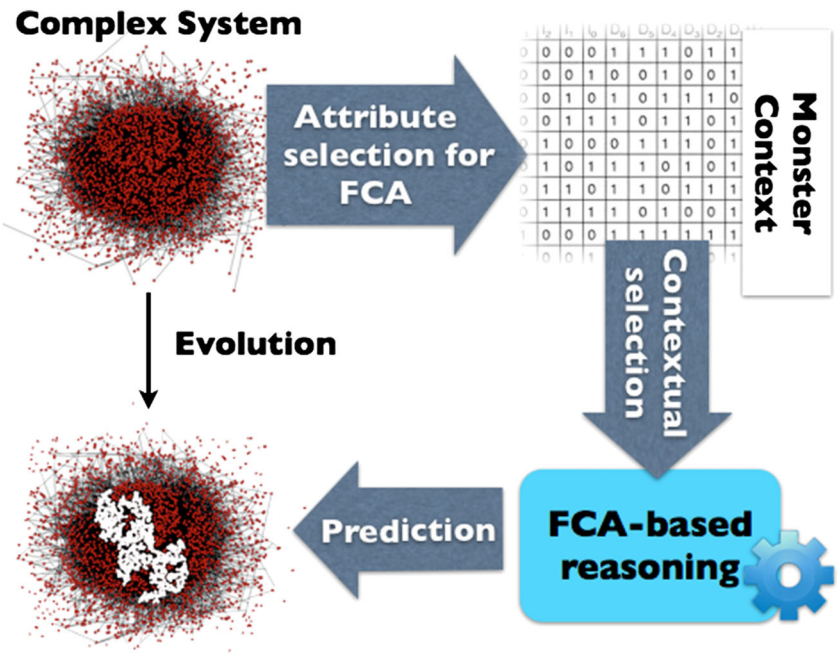

Figure 3 Formal concept analysis (FCA)-based model for qualitative reasoning with complex systems

emergent concepts from these interactions. Here, FCA can play a relevant role. In Aranda-Corral *et al.*, 2011b, this approach was applied for reasoning with contextual selections.

The full model (described in Aranda-Corral *et al.*, 2011a) is composed of events (objects), which have a number of properties (attributes). They constitute a *universal formal context* $\mathbb{M}$ (the *monster context*). Thus, $\mathbb{M}$ can be considered as the *global memory* from which subcontexts are extracted. Once the specific context is considered, it is interesting to consider the knowledge extracted from the formal context (implication basis or association rules Aranda-Corral *et al.*, 2011a).

Using Formal Concept Analysis Reasoning for Prediction

The case of inferring properties about future events when the monster model presents attributes of past events (used in forecasting, Forecasting/Prediction section) is particularly interesting. When some attributes are known to be satisfied by a future event, the inference process consists of three steps:

- The question on whether a new event (object) has a property (attribute) is raised. Some other properties (attributes) of this object are known $\{A_1, \ldots A_n\}$.
- The subcontext induced by a selection of attributes is used to compute a KB $\mathcal{L}$, called *contextual KB*. This KB consists of a set of implications among attributes, extracted from the subcontext.
- A reasoning system (Aranda-Corral *et al.*, 2011b) is executed on the contextual KB, taking $\{A_1, \ldots A_n\}$ as initial facts. The results are attributes inferred from the object.

Note that it only computes those attributes entailed from the set of attributes selected by the user. Therefore, we would need to understand the topology of the lattice to properly choose the attributes to reason with.

SCALE-FREE RESIDUE OF CONCEPT LATTICES

Given an attribute set for the objects of a CS, the CLCS is the CL built from the monster context

associated to the CS. Note that the CLCS can be considered as a directed graph (as the Hasse diagram indicates) or as a non-directed graph if necessary. The analysis of CLCS reveals interesting concepts for better understanding the structure and dynamics of CS. It is also useful to consider the role some attributes play in the qualitative study of CS (Aranda-Corral *et al.*, 2013). It should be noted that the CLCS is a complex network of semantic relationships that is not bounded by the self language, as in other semantic networks (Motter *et al.*, 2002). This is because there are concepts that are not represented by a single language term nor a intelligible definition by the observer. Complex networks with extreme structural topology are expected to appear. The complexity of such CLCS lies in the fact that the combinatorial nature of FCA covers every formal concept.

Recall that a scale-free network is one whose degree distribution follows a power law, at least asymptotically: the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim ck^{\gamma}$ where c is a normalization constant and γ is a parameter whose value is typically in the range $2 < \gamma < 3$, although occasionally it may lie outside these bounds (as we will see later). The asymptotic behavior means that, in practice, few empirical phenomena obey the power law distribution for all the values (Clauset *et al.*, 2009). It is more common for this behavior to appear from a certain threshold $x_{\min}$. The *scale-free* residue of a CLCS is the set of its nodes whose degree is greater than $x_{\min}$.

It is expected that the topological analysis of the dynamics of CLCS shows a big picture of the CS itself. Normally one will expect that the CLCS has a topology similar to other CLs, even similar to CLs associated to random formal contexts (with similar density). However, as it will be discussed in the following examples, the degree distribution is not usually very large although there are many attributes in play. As a consequence of this, the lattice is very complex, exhibiting a different topological structure than, for example, the lattices extracted from random contexts.

It is possible to refine the choice of $x_{\min}$ (Clauset *et al.*, 2009). We use $x_{\min}$ as the degree value with the maximum frequency within the CLCS (the maximum of the degree distribution). Lastly, it should be noted that CLCS are not random networks, whose degree distribution follows a Poisson law (Albert and Barabási, 2002).

The Scale-Free Conceptualization Hypothesis

The analysis of the topology of CLs is a promising method for addressing the issue raised in the introduction, namely whether sound qualitative modelizations (in our case, the CLs) share a similar structure. The rest of the paper is devoted to studying the following hypothesis, which provides a solution in a number of CS of different natures:

> **Scale-Free Conceptualization Hypothesis (SFCH):**
> Only if the attribute set selected to observe the Complex System is computable, objective, and induces a Concept Lattice that provides a sound analysis of the CS (from the point of view of some type of BR), then its degree-distribution is scale-free.

FORECASTING/PREDICTION

Two very different CS have been selected, both in which prediction (or diagnosis) is a relevant aim for humans. The first one (the Spanish soccer league) is a CS where a great number of levels, factors and agents take part. The goal is to simulate human forecasting of soccer matches (Aranda-Corral *et al.*, 2013; see previous

Table 1 Features of the complex system (CS) studied for prediction

CS	Information	Spatial feature	Context size	Information quality
Sport	Complete	No	Big	Rich
Darfur	Incomplete	Yes	Medium	Poor

Aranda-Corral *et al*., 2011a; Aranda-Corral *et al*., 2011; Aranda-Corral *et al*., 2011b). The second one is an experimental application of a CS where spatial features are relevant, the Darfur conflict (Table 1). In Aranda-Corral *et al*., 2013, the authors show that in the case of prediction of soccer bets, simple statistical forecasting rules, which are usually simplified models, produce better predictions than more complex methods. This is especially true when the future values of a criterion are highly uncertain (as it was already shown in other cases Andersson *et al*., 2003).

Complex Systems in Sport

The first case represents a nice example of successful application of BR on CS by means of FCA tools. The study of CS associated to sport is increasing in importance due to economic (and political) reasons. Several pieces of evidence about prediction in betting markets or match forecasting have encouraged many projects in this field of application. In Aranda-Corral *et al*., 2013, we focus our efforts on forecasting of soccer results. In this case study, a monster model is constructed by considering matches as objects and selecting qualitative attributes related to team features and previous results (Aranda-Corral *et al*., 2011). The monster context covers—among other information—all the matches from the 2005/06 season to the 2010/11season of the Spanish Premier League. The attributes express properties of the teams involved in the match, and they are booleanized if necessary (Aranda-Corral *et al*., 2011a; Aranda-Corral *et al*., 2011).

Table 2 shows the main parameters of the cumulative CLCS. Figure 4 shows the log-log

Table 2 Data on accumulated concept lattice associated to complex systems for soccer (up) and Darfur (down)

	$\lvert\mathbb{O}\rvert$	$\lvert\mathbb{A}\rvert$	Density (%)	$\lvert CL\rvert$	$< k >$	$x_{\min}$	*sc* (%)	γ
05/06	842	94	13.31	27 434	8.27	7	77.62	5.33
05/07	1684	94	13.30	81 490	9.47	9	60.87	6.08
05/08	2526	94	13.26	140 739	9.97	9	68.2	6.34
05/09	3368	94	13.37	243 959	10.62	10	63.6	6.84
05/10	4210	94	13.36	324 146	10.82	10	66.20	6.69
3-month intervals	$\lvert\mathbb{O}\rvert$	$\lvert\mathbb{A}\rvert$	Density (%)	$\lvert CL\rvert$	$< k >$			
1	787	81	14.73	294		5.310		
4	3148	81	14.63	5864		7.676		
7	5509	81	13.63	44 911		9.595		
10	7870	81	10.61	62 870		9.678		

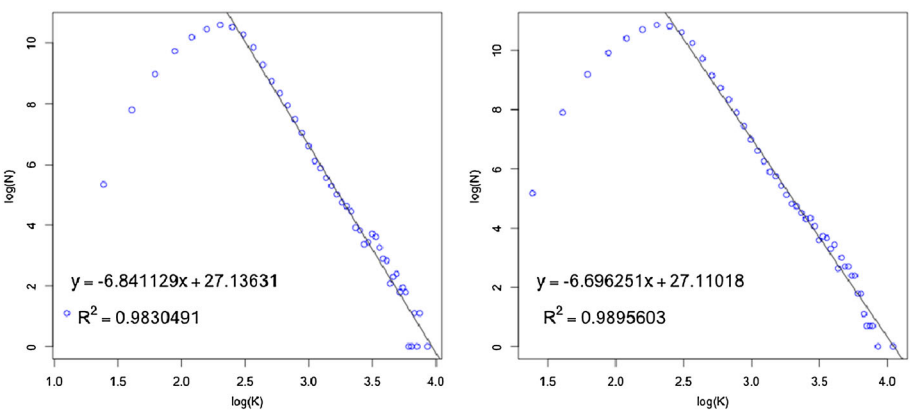

Figure 4 Log-log degree distribution of concept lattice associated to complex systems for sport forecasting (four and five seasons)

scale distribution that represents the scale-free residue behavior.

It is worthwhile to note the growth and size of γ in the distributions shown in Table 2. One explanation for this behavior is based on Barabási *et al.*, 1999, where the asymptotic behavior of scale-free networks is studied. In the work mentioned, two models of limit behavior are shown. In one of them, the number of nodes in the network remains constant, and the distribution tends to be a Gaussian one. In our case, if it is supposed that the process starts from a node set with 2^{94} nodes (the power set of attribute set), and in each step, new objects are added, new relations in the lattice emerge. From this point of view, the CLCS shows a distribution that is between pure scale-free and Gaussian (in the sense described in Albert and Barabási, 2002; Barabási and Albert, 1999; Barabási *et al.*, 1999). According to SFCH, the closer the CLCS degree distribution is to being purely scale-free, the richer and more useful the information on CS dynamics will be.

It is not possible to predict how the attribute distribution of new objects (matches) will be; therefore, only the complexity of the reasoning process (production system execution) of their initial attributes can be estimated. In order to figure out how the frequent rule-based computations will be (to better understand the complexity of logical reasoning in a CLCS), it is interesting to analyse the topological structure of concept nodes that are involved in frequent deduction tasks. For example, Figure 5 (left) depicts *deduction network of level 10* of CLCS for sport forecasting.

The deduction network of level p is built as follows. Let $RL_{\max}$ be the maximum number of rules whose left side is contained by the extension of one concept (meaning, this value is set by the concept whose intension contains the left side of more rules). In the same way as for $RL_{\max}$, $RR_{\max}$ is computed, this time considering, the right side of the rules. The nodes of the deduction network are the concepts of CLCS whose intents contain both the left side of at least the p% of $RL_{\max}$ rules, and the right side of at least p% of $RR_{\max}$ rules in the KB extracted from the context.

For example, the deduction network of level 10 contains 715 nodes and 1296 arcs and its diameter is 7. The diameter (considered as a directed graph) suggests that a reasonable (upper bound) estimation of rule firings by the production system is 7 (a deduction made by the reasoning system induces

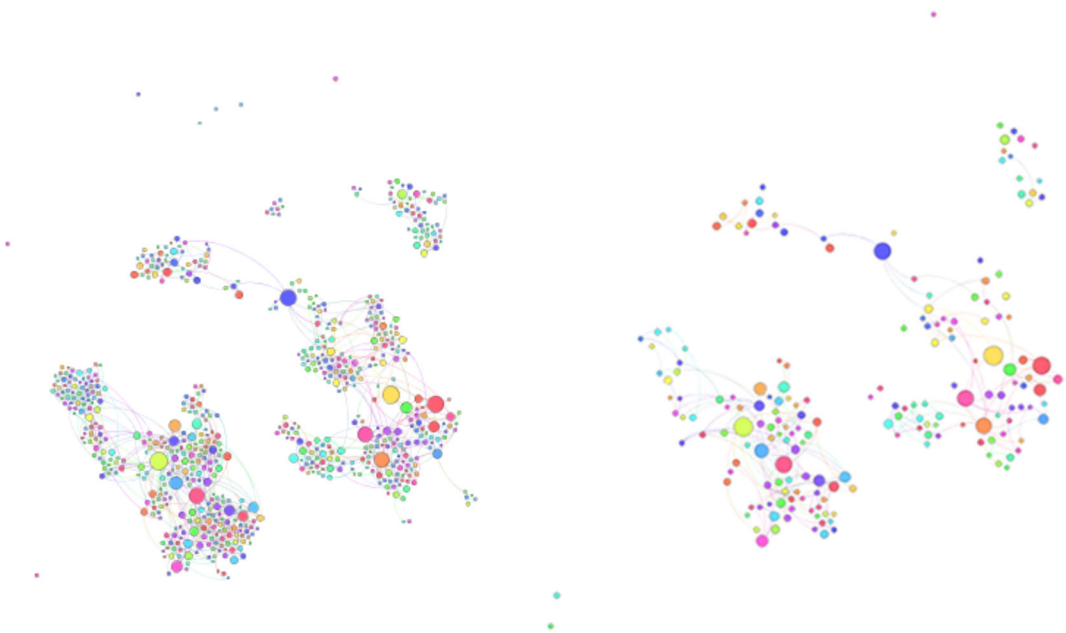

Figure 5 Deduction graph for sport forecasting for $p = 10$ (left) and $p = 20$ (right)

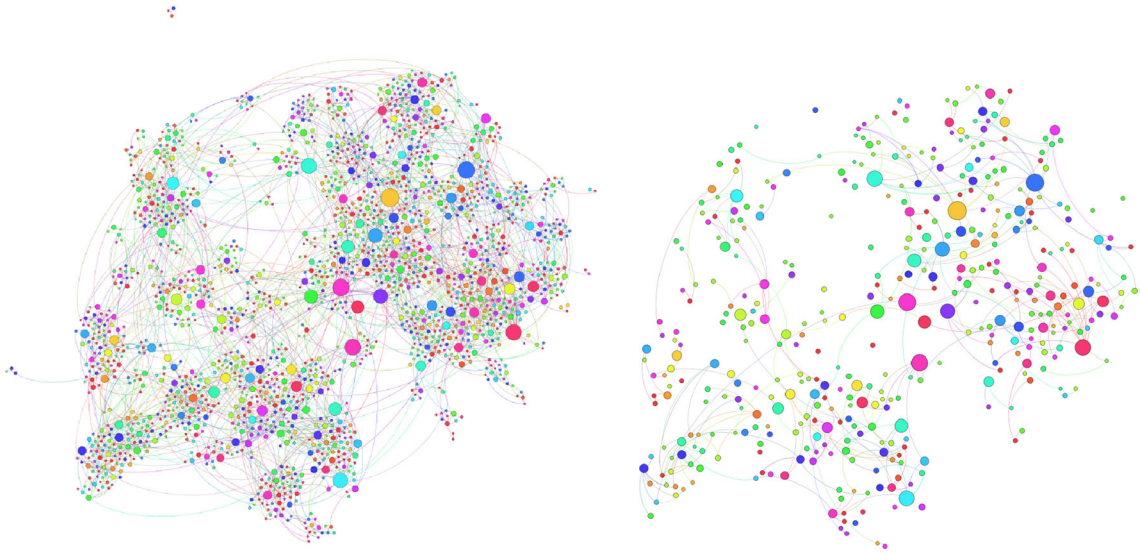

Figure 6 Deduction graph for the alternative attribute set on sport forecasting for $p = 10$ (left) and $p = 20$ (right)

a path in the CLCS). The existence of many connected components is useful to decide which attributes are relevant to frequently achieve a certain result by prediction (those that appear in the same connected component as the attributes related with the goal). The deduction network keeps the same properties for other values of p (see Figure 5 right for $p = 20$; Figures 6).

Darfur Conflict

The second CS is the Darfur Conflict. The importance and its character as a contemporary, relatively unknown episode of genocide are the main motivations for working with the information available about this topic. The conflict represents a very CS, and the information about its dynamics can be collected from different sources (the US government, non-governmental organizations, academic research groups, etc.).

The key feature that differentiates the multilevel nature of this CS from the one aforementioned is that the information is more difficult to process and to compute than in the former example. Excellent references such as Hagan and Rymond-Richmond (2008) provide a number of insights on the conflict that may be considered. Furthermore, geolocation services allow the consideration of spatial features of events and interactions in the region. It is worth noting that the information increases in a different way that is in the CS of sport forecasting: via multiple resources, heterogeneous data and frequent incorporation of new information resources. In the current state of the project, the analysis uses the already cited Hagan and Rymond-Richmond (2008) and other sources: *aerial military attacks on civilians and humanitarian agents in Sudan, 1999–2011*,[2] *Darfur - Destruction of 1,000 Villages*,[3] *Crisis in Darfur*[4] and *North Sudan*.[5] Nevertheless, a number of new information sources and data will be considered shortly.

Preliminary experiments with this data showed a significant number of false positives (meaning that the system predicts attacks to villages that did not really occur) and a small number of false negatives (something more serious). This behavior could be explained by two reasons. On one hand, the information contained in the data is poor in some cases (there even exist a number of ambiguities), affecting the variety and accuracy of the attributes. The

[2] http://www.sudanbombing.org/
[3] http://bbs.keyhole.com/ubb/ubbthreads.php?ubb=showflatNumber=721111site_id=1
[4] http://www.ushmm.org/maps/projects/darfur/
[5] http://bbs.keyhole.com/ubb/ubbthreads.php?ubb=showflatNumber=393317

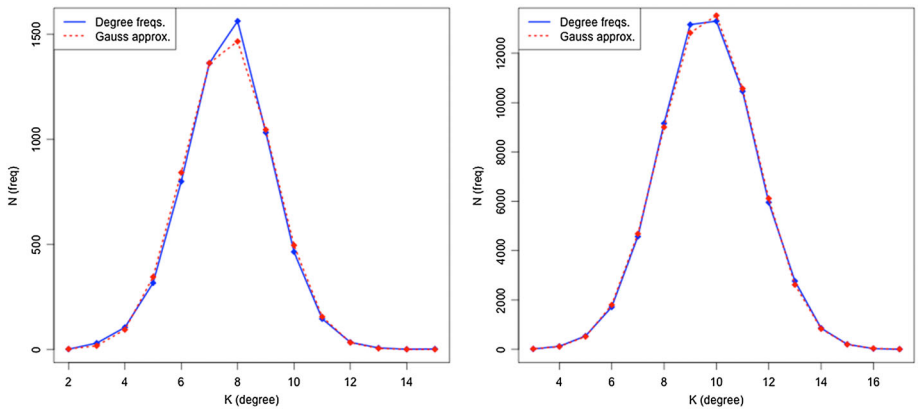

Figure 7 Degree distributions of Darfur complex system (4 and 10 3-month periods) and their Gauss approximations

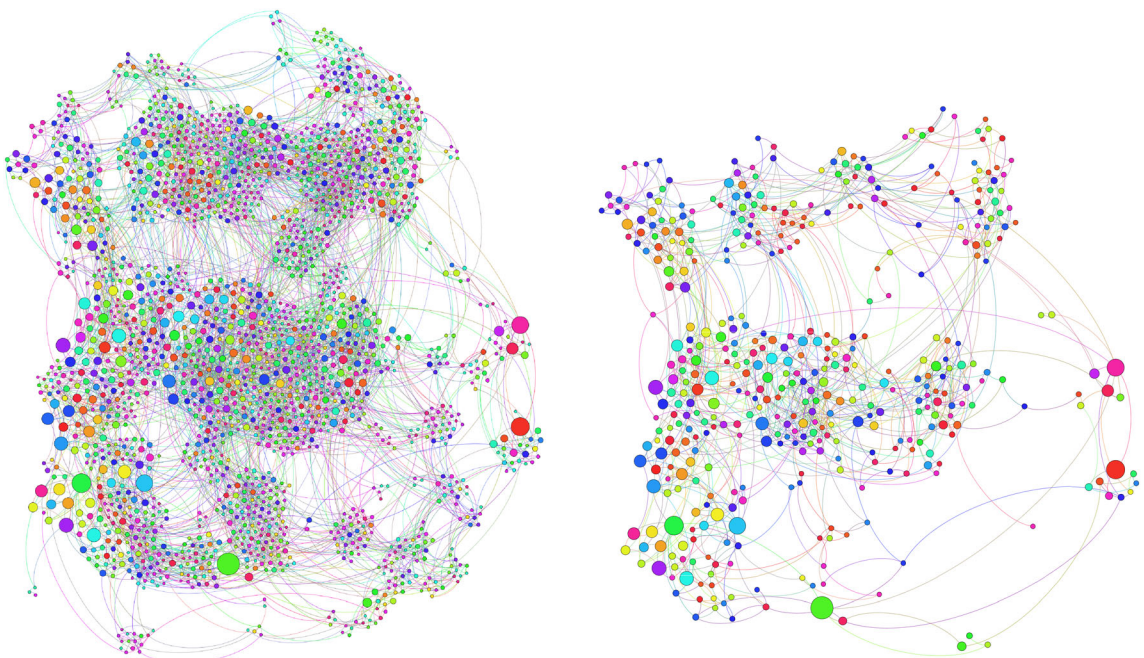

Figure 8 Deduction graph for Darfur $p = 10$ (left) and $p = 20$ (right)

Gauss distribution of node degrees in the associated CLCS (Table 2 and Figure 7) indicates that there are no key features within the concept population to exploit in the reasoning process. This fact is reinforced when it is found that the cumulative CLCS stabilizes after 10 periods (that is, there are no any new relevant information). This phenomenon suggests that to obtain a scale-free distribution, more attributes are necessary to increase the number of concepts, which is one of the requirements shown in Albert and Barabási (2002) to asymptotically obtain a free-scale network.

On the other hand, from the point of view of the reasoning, the deduction graphs of Levels 10 and 20 for the CLCS show a very complex network (Figure 8), where only one period of the conflict is represented (1986 nodes and 6413 links). It is highly connected, suggesting that it does not provide any information about which attributes are more relevant to predict a certain result.

Finally, it is interesting to remark that target attributes (the concept 'attack a village') in CLCS associated to Darfur are not in the tail of the distribution, whereas in sport forecasting, target attributes (win, lose and drawn) have a high degree. This fact can explain how the attribute selection is essential in this last case.

Testing Scale-Free Conceptualization Hypothesis with Other Concept Lattices: Attribute Set Influence

The quality of available data on a domain is very sensitive when reasoning techniques are applied to it. In the former section, the correlation between the quality of available data and the SFCH was demonstrated. In this section, it will be shown that the quality of available information is as relevant as the choice of the attribute set used to model the problem.

In the domain of soccer results, most of the parameters considered are quantitative. Thus, one of the most sensitive parts of the attribute creation/modelization process is the choice of proper thresholds to be used in the data discretization process.

In a second experiment, a different attribute set was built to model the CS in sports results forecasting. Most of the attributes present in this new attribute set were also present in the former one, but had different thresholds. The former attribute set tried to represent/capture the regular behavior of teams in competition to be able to predict results.

The changes in this new attribute set affected not only the performance of predictions produced by the system (which decreased significantly) but also to the degree distribution of the CLCS (Figure 9 and Table 3).

To test the performance of predictions, two attribute selections on the monster context have been considered, one taken from the new attribute set and one from the former. Both are of the same size

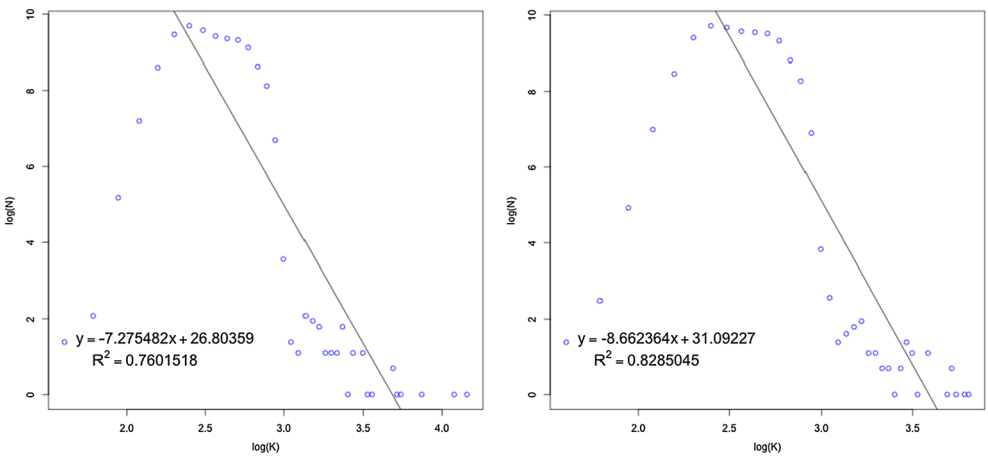

Figure 9 Log-log degree distribution of concept lattice associated to complex systems for the alternative attribute set of sport forecasting (four and five seasons)

Table 3 Data on accumulated concept lattice associated to complex systems for the alternative attribute set on soccer

	$\|\mathbb{O}\|$	$\|\mathbb{A}\|$	Density (%)	$\|CL\|$	$< k >$
05/06	842	26	50.51	46 439	11.39
05/07	1684	26	50.76	75 032	12.26
05/08	2526	26	50.88	91 625	12.64
05/09	3368	26	50.83	104 261	12.92
05/10	4210	26	50.91	115 206	13.15

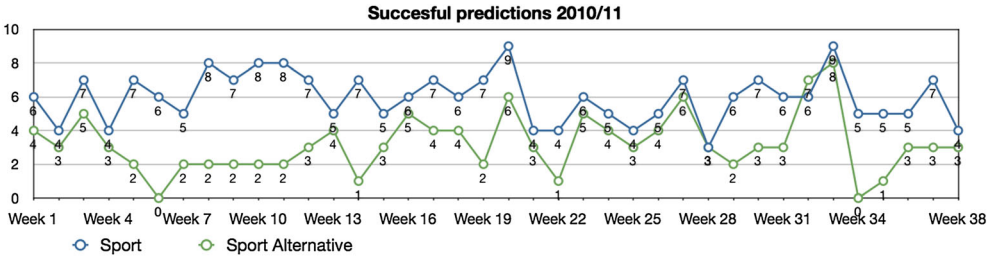

Figure 10 Successful predictions for season 2011 using two different attributes selection

Table 4 Data on concept lattice associated to complex systems for other contexts (see also Figure 11)

	$\lvert\mathbb{O}\rvert$	$\lvert\mathbb{A}\rvert$	Density (%)	$\lvert CL\rvert$	$< k >$	$x_{\min}$	*sc* (%)
M	8124	119	19.33	238 709	11.453	11	63.04
T^3	958	29	34.48	59 504	10.637	9	67.27
W	68	178	20.48	25 408	9.370	8	70.92
C	1728	25	28	12 639	9.438	7	98.96

(12 attributes). Successful predictions for the 2011 season are shown in Figure 10.

Finally, the deduction graph of level 10 in this case (Figure 6) shows a less complex (1899 nodes and 3480 links) network than in the case of Darfur, but one that is more complex than that of the soccer case.

CLASSIFICATION

To illustrate the SFCH in the descriptive analysis of other systems (oriented in this case for classification of objects within the CS), four contexts have been selected (see Table 4):

Mushrooms[6] (context *M*): A mushroom dataset, being described in terms of physical characteristics, where objects are mushrooms and attributes are qualitative properties of the mushrooms. The formal context associated to the data can be understood as a qualitative description of the current state of evolution of a type of vegetables.

Tic-tac-toe[7] (or T^3): End results of the tic-tac-toe game. The objects are possible results of the game, and the attributes describe the configuration of the board at the end of the game. This dataset has been selected as a non-CS example, to obtain a CL from a well-known formal context that does not come from a CS.

Wine[8] (context *W*): A dataset containing different wines (objects) where the attributes are their qualitative properties. It could be enriched with more specific attributes related to biological systems associated with its production/processing (Borneman *et al*., 2009).

Car evaluation[9] (context *C*): A dataset representing the acceptability of cars. The objects are cars, and the attributes are the subjective qualitative properties of cars. For the purposes of this paper, the context collects qualitative information about the automobile industry, its evolution and products.

Two of those contexts, the wine and mushroom contexts, consist of a cataloguing carried out by humans of the result of the evolution of a CS. Those show a scale-free residue ($R^2 > 0.95$). The context T^3 is not based on classification of a CS, and the distribution satisfies $R^2 < 0.75$. Lastly there is a special case, the car evaluation context, which displays an intermediate behavior that we conjecture to be a consequence of the attempt to describe objects

[6] http://archive.ics.uci.edu/ml/datasets/Mushroom
[7] http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame
[8] http://archive.ics.uci.edu/ml/datasets/Wine
[9] http://archive.ics.uci.edu/ml/datasets/Car+Evaluation

produced within a CS (the automobile industrial complex) by means of subjective attributes (opinion), thus allowing for contradictory opinions. This borderline case study presents an intermediate degree distribution (see Figure 11) with $0.85 < R^2 < 0.9$

WORDNET SUBSYSTEMS

As it was already mentioned, FCA provides concepts (semantically sound attribute sets), which may not be definable in terms of the self attribute language. These languages (attribute sets) are ad hoc, but it is unknown what occurs if we choose an existing (real) language. Under SFCH, it should occur that real languages provide sound cognitive term sets semantically organized by scale-free law. To demonstrate this, the *WordNet* database (http://wordnet.princeton.edu/) was considered. WordNet is a lexical database of English in which words are grouped into sets of cognitive synonyms (synsets). WordNet represents the state-of-art of our own language evolution in such a dimension. To build the formal context, each single word has been considered as an object and each possible synset as an attribute, in which an object owns an attribute if the word belongs to the corresponding synset. WordNet, considered as a formal context, produces a CL very similar to the structure of the synsets. Two smaller word subsets have been considered (Table 5): adverbs and verbs. In both cases SFCH holds (Figure 12).

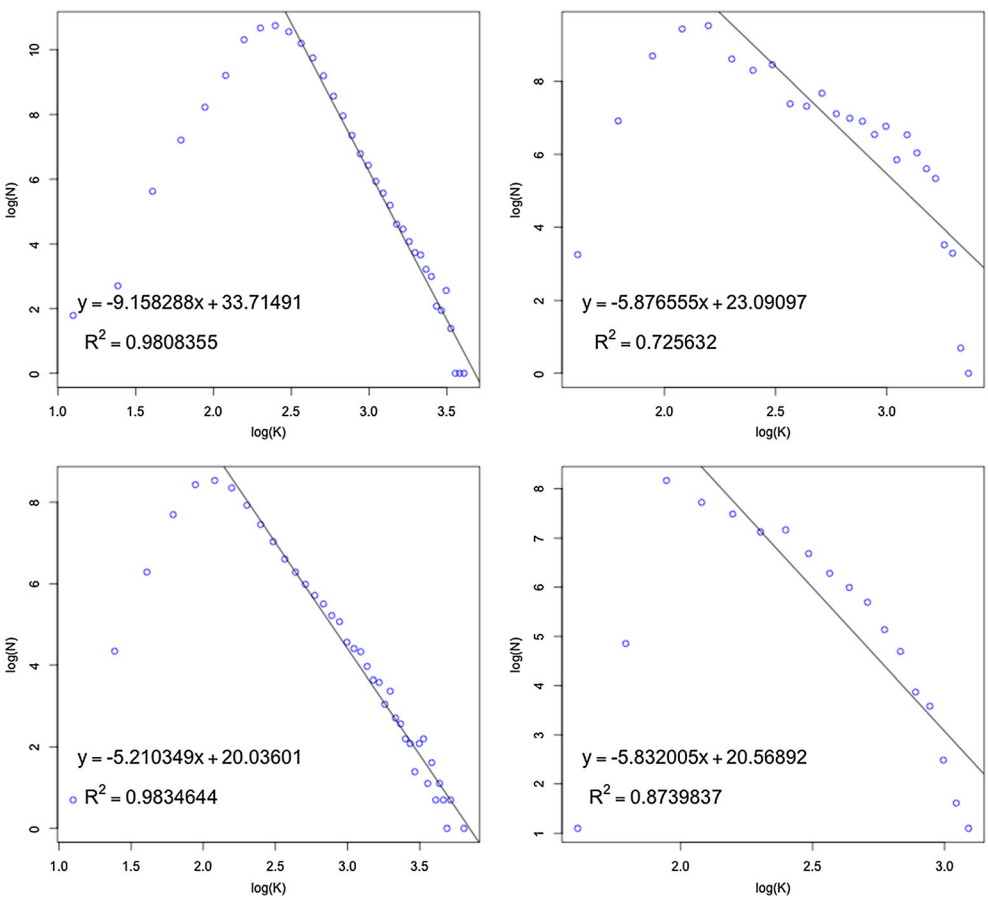

Figure 11 Log-log degree distribution for mushroom and T^3 (up), wine and car evaluation contexts (down)

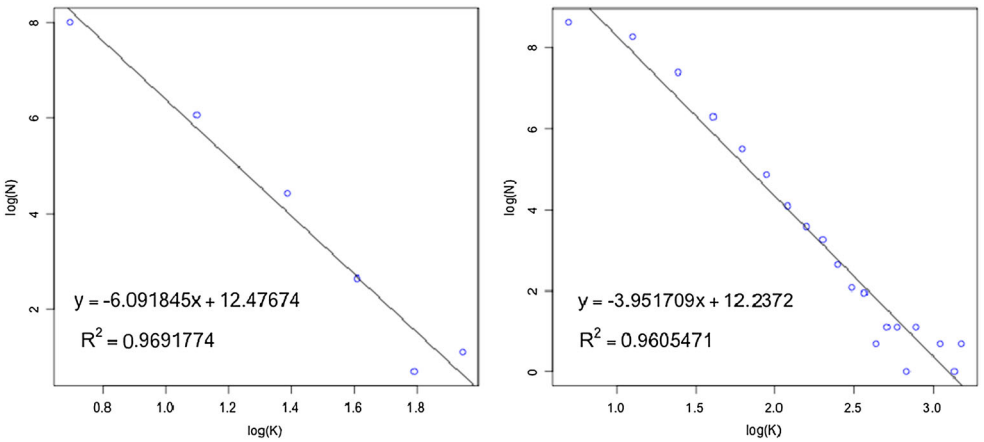

Figure 12 Log-log degree distribution for adverbs and verbs subcontexts of WordNet

Table 5 Scale-free conceptualization hypothesis for WordNet subsets

	$\|\mathbb{O}\|$	$\|\mathbb{A}\|$	Density (%)	$\|CL\|$	$< k >$	*sc* (%)
Adverbs	4481	3621	0.03	3529	2.187	100
Verbs	11 529	13 767	0.02	12 222	2.967	100

RANDOM CONTEXTS

In order to show how specific the topological structure of CLCS is, we compare it with the CL associated to random contexts (that is, random Boolean matrices). Such contexts represent a set of observations (objects with attributes) of a system whose behavior seems to be random, from the perspective of the selected attribute set. It is expected that the absence of a strong relationship among attributes will be represented in the associated CL. Random contexts (and associated CL) are useful to contrast the SFCH. To show whether a scale-free residue exists or not in a certain type of CLCS, an experiment was carried out to compare the CLCS with the CLs associated to random contexts. In this experiment, two sets of 10 000 random formal contexts with a fixed density, number of objects and number of attributes were generated (Table 6).

The fixed parameters take the size and dimension values of the first two monster contexts for soccer (Table 2). Results show that the degree distributions of CLCS are very different from the distributions of CL associated to random formal contexts (a random context can be interpreted as the result of the observation of the CS by means of non-relevant attributes or a qualitative observation of a chaotic system). In Figure 13, the degree distribution of CL associated to random formal contexts is shown, which does not follow a power law distribution. Also this has

Table 6 Data on random formal contexts and its concept lattices. The parameters $|\mathbb{O}|$, $|\mathbb{A}|$ and density are taken from the formal contexts on soccer domain

Dataset	$\|\mathbb{O}\|$	$\|\mathbb{A}\|$	Density (%)	Average R^2
m1	842	94	13.31	0.847606
m2	1684	94	13.30	0.8037136
m3	2526	94	13.26	0.809734

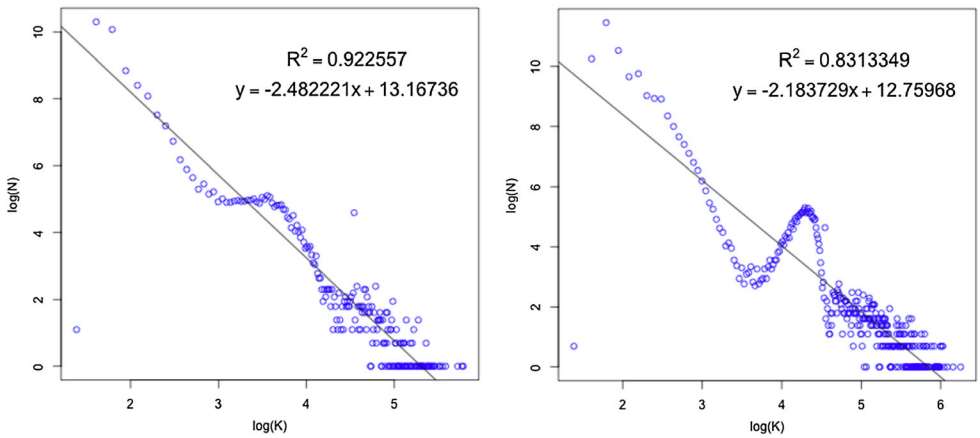

Figure 13 Two representative examples of log-log degree distribution of random contexts with density similar to sport forecasting (accumulated for one and two seasons)

been studied by means of a goodness of fit test (χ^2 test), to prove that the degree distribution of CL associated to random formal contexts does not follow the power law distribution from $x_{\min}$. Finally, it is worth noting that in the degree distribution of CL associated to random formal contexts, a kind of *phase transition* always appears (see also Figure 13).

CONCLUDING REMARKS AND FUTURE WORK

This paper is devoted to studying the relationship between the structures of qualitative representations of CS and their usefulness in solving basic problems of CS research, such as prediction, classification and descriptional complexity. In order to clarify the results, only classical FCA is used (that is, Boolean attributes or booleanization of attributes by means of thresholds), but it could be expanded to multivalued attributes by using other well-known methods developed for this kind of attribute (Ganter and Wille, 1999).

Throughout the paper, a number of CLCS have been studied. Table 7 shows their main features. It is worthy to note that, to cover a wide variety of cases, different kinds of BR have been considered, both within the CS and in its analysis. The first is the one used for modelling the system (extracting a set of recognizable qualitative features for describing phenomena/objects), BR1. The second one represents the use of BR by the self agents, which lives within the system to study, BR2. The third use of BR, BR3, considers that the designed system uses BR to simulate agents within the CS to achieve the goal. The last one is used only in a specific case, where the aim is to simulate a specific behavior within a CS.

The SFCH relates CLCS topology with the information about CS in two interesting CS, (*Sport* and *Darfur conflict*). The positive case is thoroughly analysed in (Aranda-Corral *et al.*, 2013). The negative one, the case of the Darfur conflict (with Gauss distribution) shows that the use of data of poor quality produces a CLCS with a topology that prevents successful reasoning, because it does not allow sound discrimination of target attributes. We are currently adding new attributes and data to the last CS in order to enrich the information. A third example, using another attribute set in the domain of *Sport* shows that a bad attribute selection can be as bad as the use of poor quality data.

Concept lattice associated to CS strongly depends on the use of specific terms (attributes), as they have a direct influence on concept relationships. Because in CL, each concept represents the definition of a new term; CLCS should be similar to semantic networks. In Motter *et al.*, 2002, the semantic network among the concepts expressed by (English language) terms is studied, which also has a scale-free distribution. In

Table 7 Description of the examples used in the paper

System	CS?	Aim	Objective Data?	Accuracy of data?	BR-Modelling (BR1)	BR-agents within CS (BR2)	Modelling reasoning of agents within CS (BR3)	Scale Free?	SFCH?
Sport	Yes	Human bet simulation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Darfur	Yes	Prediction	Yes	Poor	Yes	Yes	No	No	Yes
Mushroom	Yes	Classification	Yes	Yes	Yes	No	No	Yes	Yes
T^3	No	Classification	Yes	Yes	No	No	No	No	Yes
Wine	Yes	Classification	Yes	Yes	Yes	Yes	No	Yes	Yes
Car	No	Evaluation	Medium	No	Yes	Yes	No	No	Yes

CS, Complex System; BR, bounded reasoning; SFCH, scale-free conceptualization hypothesis.

contrast to the aforementioned work, our network is not sparse, because we do not use terms, but rather concepts (which are specified by an attribute set and not by an isolated term). In Motter *et al*., 2002, $P(k) \sim ck^{\gamma}$, with $\gamma = 3.5$, but in the distribution of CLCS studied here $\gamma > 3.5$, suggesting that the scale-free residue contains a very small amount of nodes with high connectivity. The network representing words and synonymic relationship between them analysed in Albert and Barabási (2002) has $\gamma = 2.8$. Another example is the semantic network associated to Roget's Thesaurus (Roget, n.d.; Steyvers and Tenenbaum, 2005), with $\gamma = 3.19$. It remains to be studied what happens if we expand the language by inserting new language terms defined from the available information.

In another experiment, two more CLCS of forecasting of soccer results have been analysed. Those obtained from the two attribute selections used to obtain the predictions that are shown in Figure 10. The aim was to investigate if the SFCH could also be useful to refine the attribute selection taken (from the monster context) to reason about the CS. Although the first experiments showed promising results, a deep analysis of the relationships between reasoning entailment and CLCS need to be performed.

As preliminary results, it is worth noting that both attributes selections respect the SFCH (Figure 14). This means that the CLCS associated to the attribute selection that produces good predictions presents a scale-free degree distribution, whereas the other does not.

Another interesting feature observed is the size of the CL associated to both attribute selections. Whereas the attribute selection that produces good predictions has 198 concepts, the other has 3094. Taking into account that the number of attributes and the object set considered are the same, it suggests that a simpler CL will provide sounder reasoning as was already observed in the analysis of Forecasting/Prediction section. A more thorough analysis of the internal structure of these CLs could lead to clues on how to modify the current attribute selection in order to improve it.

Future work will also focus on finding a relationship between the human involvement in CS

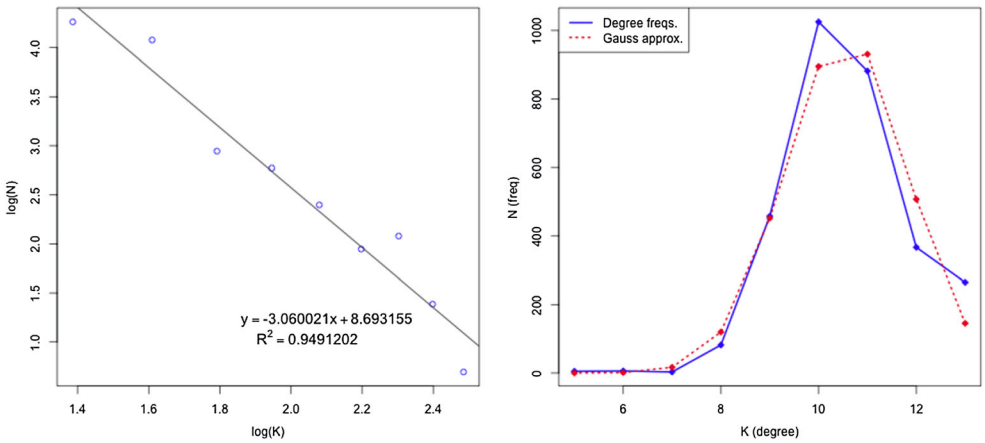

Figure 14 Log-log degree distribution degree distributions for well selected attributes concept lattice (right) and degree distribution and its Gauss approximation for the alternative one (left)

and the degree distribution of CLCS, in order to discover how useful the information behind CS could be in making short-term predictions. Although the experiments show that SFCH is a sound work hypothesis, it is only descriptive. We aim to demonstrate the strong ties among SFCH and the performance of the reasoning systems designed on the CLCS.

ACKNOWLEDGEMENTS

Supported by TIN2009-09492 project of Spanish Ministry of Science and Innovation, and *Excellence project* TIC-6064 of *Junta de Andalucía* cofinanced with FEDER founds.

REFERENCES

Albert R, Barabási A-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**: 47–97.

Andersson P, Ekman M, Edman J. 2003. Forecasting the fast and frugal way: a study of performance and information-processing strategies of experts and non-experts when predicting the World Cup 2002 in soccer. Working Paper Series in Business Administration. Stockholm School of Economics, 9.

Aranda-Corral GA, Borrego-Díaz J, J Giráldez-Cru. 2012. Agent-mediated shared conceptualizations in tagging services. Multimedia Tools and Applications.

Aranda-Corral GA, Borrego-Díaz J, Galán J. 2011a. Bounded rationality for data reasoning based on formal concept analysis. Proc. DEXA 350–358.

Aranda-Corral GA, Borrego-Díaz J, Galán J. 2011b. Confidence-based reasoning with local temporal formal contexts. Proc. 11th Int. Conf. Artif. Neural Networks, LNCS vol. 6692. Springer, 461–468.

Aranda-Corral GA, Borrego-Díaz J, Galán J. 2013. Complex concept lattices for simulating human prediction in sport. *Journal of Systems Science and Complexity* **26**(1): 117–136.

Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J. 2011. Selecting attributes for sport forecasting using formal concept analysis. ECAL, Workshop on Complex Systems in Sports.

Aranda-Corral GA, Borrego-Díaz J. 2010. Reconciling knowledge in social tagging Web services. Proc. 5th Int. Conf. Hybrid AI Systems (HAIS 2010), Lecture Notes in Artificial Intelligence, vol. 6077. Springer, 383–390.

Balcázar JL. 2010. Redundancy, deduction schemes, and minimum-size bases for association rules. *Logical Methods in Computer Science* **6**(2): 1–23.

Barabási AL, Albert R. 1999. Emergence of scaling in random networks. *Science* **286**: 509–512.

Barabási A-L, Albert R, Jeong H. 1999. Mean-field theory for scale-free random networks. *Physica A* **272**: 173–187.

Borneman AR, Chambers PJ, Pretorius IS. 2009. Systems biology as a platform for wine yeast strain development in biology of microorganisms on grapes. In Must and in Wine, König H, *et al*. (eds.). Springer: Berlin-Heidelberg, Germany; 395–414.

Bourgine P, Chavalarias D, Perrier E (eds.). 2009. The CSS roadmap for the science of complex systems. http://www.assystcomplexity.eu/db/assyst/ASSYST-roadmap2009-2.pdf

Clauset A, Shalizi CR, Newman MEJ. 2009. Power-law distributions in empirical data. *SIAM Review* **51**(4): 661–703.

Ganter B, Wille R. 1999. Formal Concept Analysis - Mathematical Foundations. Springer: Berlin-Heidelberg, Germany.

Goldstein DG, Gigerenzer G. 2002. Models of ecological rationality: the recognition heuristic. *Psychological review* **109**(1): 75–90.

Guigues J-L, Duquenne V. 1986. Familles minimales d' implications informatives resultant d'un tableau de donnees binaires. *Mathématiques et Sciences Humaines* **95**: 5–18.

Hagan J, Rymond-Richmond W. 2008. Darfur and the Crime of Genocide. Cambridge University Press: Cambridge, UK.

Luxenburger M. 1991. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines* **113**: 35–55.

Motter AE, de Moura APS, Lai Y, Dasgupta P. 2002. Topology of the conceptual network of language. *Physical Review E* **65**: 065102(R).

Roget PM. Roget's Thesaurus of English words and phrases (1911 ed.). Retrieved June 21, 2012, http://www.gutenberg.org/etext/1068

Simon HA. 1982. Models of Bounded Rationality. MIT Press: Cambridge, MA.

Steyvers M, Tenenbaum JB. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* **29**(1): 41–78.