STUDIA UBB PHYSICA, LVI, 2, 2011

CORRELATION CLUSTERING APPROACH TO LOGICAL LEARNING

CSEH GY[1], NÉDA Z.[1,2], D. DAVID[3,4]

ABSTRACT. The relation between the level of logical learning and intelligence quotient (IQ) of an individual is investigated in view of a complex clustering model. A useful analogy between the learning process of a material with logically interconnected parts and the correlation clustering problem suggest a phase-transition like trend: the Logical Learning Rate of individuals should increase sharply above some IQ level. Psychological tests are conducted and analyzed for a sample of 177 high-school students. The measured IQ and LLR values confirm the conjecture.

Keywords: *correlation clustering, phase-transition, logical learning, networks*

INTRODUCTION

Basic models and methods of statistical physics proved to be useful in understanding several complex phenomena in biology, economy or sociology [1]. Critical phenomena, pattern formation and selection, self-organized criticality, small-world networks, nonlinear and disordered systems, non-equilibrium processes and combinatorial optimizations are only a few known examples in such sense. In the present work we consider the recently introduced correlation clustering problem [2-5] and an interesting application of it in psychology.

Correlation Clustering (CC) can be formulated simply in laymen terms: given a set of elements interconnected globally through positive and negative links, find an optimal clustering of them which maximizes the number of positive links within the clusters and the number of negative links between the clusters. Based on our daily-life experience a simpler sociological formulation can be also given. Let us consider a group of people that know each other and have fixed and reciprocal (symmetric) propensities towards the group members. These propensities can be either positive (they sympathize each other) or negative ones (they consider antipathetic the other). The CC problem is to find an optimal grouping of them, so that the given propensities are optimally satisfied. This

[1] *Babeş-Bolyai University, Department of Physics, Cluj-Napoca, Romania*
[2] *Interdisciplinary Computer Simulation Research Group, KMEI, Cluj-Napoca, Romania*
[3] *Babeş-Bolyai University, Department of Clinical Psychology, Cluj-Napoca, Romania*
[4] *Mount Sinai School of Medicine, New York, USA*

means that persons connected with positive propensities (persons that “like” each other) should be in the same group while persons connected by negative propensities (agents that “hate” each other) should be in different groups.

It is easy to realize that the CC problem is relevant to many other practical situations in different domains of sciences. It was originally motivated by a research at Whizbang Labs, where learning algorithms were trained to help various clustering tasks [1]. CC is also related to agnostic learning [6], which is an emerging approach to efficient data mining and artificial intelligence. Important applications can be in medicine and pharmaceutics, where one needs to divide drugs in compatibility groups. Closely related problems were considered also while studying coalition formation phenomena in sociological systems [7-9]. It also resembles the infinite-range Potts-glass [10-12] system. From statistical physics point of view it is especially interesting because it exhibits a non-trivial phase transition-like phenomena [3].

In the present work we apply results of the correlation clustering problem for explaining experimentally obtained connections between the Logical Learning Rate (LLR) and Intelligence Quotient (IQ) of a person. The relationship between learning and intelligence is one of the major topics in psychology [13]. There are hundreds of studies investigating this relationship, showing that explicit learning (e.g., logical learning) strongly correlates with what we traditionally call intelligence quotient. Indeed, Reber et al. [14] showed that IQ accounted for almost 50% of the variance in the explicit (logical) learning performance.

The structure of the paper is as follows: first, we will review those major results for CC that are used for explaining the experimental data, than we discuss the connection between CC and logical learning and finally we present our experimental results for the relation between LLR and IQ which confirm our conjecture.

THE CORRELATION CLUSTERING PROBLEM

Solving the CC problem is a rather complex task. A perfect solution for a general situation is usually not possible because there is no grouping so that all connections are optimally satisfied. A simple example for this is when there is a “frustrated triangle”: three agents interconnected with two positive links and one negative link. If one would put all three agents in one cluster the negative link becomes frustrated, putting them in two different clusters will frustrate one of the positive links. There are of course a few simple cases when a perfect solution can be achieved. An immediate example for the three agent situation is when all propensities are positive, which means that everybody likes each other. In this case the ideal clustering is to put all agents in one cluster. Another simple case is when all propensities are negative (everybody hate each other). The solution is again straightforward: all agents have to be in separate clusters. For a general system however the problem becomes an NP hard optimization. This means that the computational time necessary for finding the optimal clustering increases as a function of the system size faster than any polynomial function. This is the reason why an exact solution by exhaustive search is impossible for moderately large systems. As an example, for a group formed by N=15 individuals and an arbitrary propensity distribution among the group members, it is not possible to obtain an exact solution in reasonable computational time even on the fastest supercomputer.

The CC problem can be formulated mathematically. For this we quantify the positive links as +1 and the negative links as -1 and introduce a K cost-function [3] which increases by 1 whenever two conflicting agents are in the same cluster or when two agents with positive propensities between them are in different clusters. The mathematical problem is then to get a clustering which minimizes the value of K:

$$K = -\sum_{i<j} \delta_{\sigma(i)\sigma(j)} J_{ij} + \frac{1}{2}\sum_{i<j} (J_{ij} + |J_{ij}|), \tag{1}$$

In Eq. 1 $\sigma(i)$ denotes the cluster to which agent i belongs, the sums are for all possible pairs, δ is the Kronecker-delta symbol ($\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$) and $J_{ij} = \pm 1$ is the link (propensity) between agent i and j. Solving the CC problem is equivalent with finding the $\sigma(i)$ values that minimizes the K cost-function. Although perspectives for a simple and exact solution are quite gloomy due to the NP hard complexity of the problem, surprisingly the solution in the thermodynamic limit (infinitely large system, $N \to \infty$) is simple! Based on analogies with thermodynamic systems and the Potts glass problem in physics, it has been shown [3-5] that for an infinitely large system, statistically the optimal clustering is the following: whenever there are more positive propensities than negative ones put all agents in one big cluster, whenever there are more negative propensities than positive ones put each agent in a separate cluster. The term, “statistically”, means here that the above solution is true for the majority of the cases, neglecting a few special situations. Before getting too excited about this, let us remember that all practically interesting cases are for finite N values, where the problem remains NP hard [15]. In such cases the best we can do is to consider some numerical optimization techniques to cluster the system. Various methods are known, the ones which have been already considered for this problem are the simulated annealing, analytical or numerical renormalization approach or a Molecular Dynamics motivated optimization trick [3-5].

The CC problem leads to a phase-transition like behavior in the thermodynamic limit [3]. In order to understand this we will introduce some relevant quantities. Let us denote by *q* the density of the positive propensities in the system:

$$q = \frac{W_+}{W_t} \quad (3)$$

(W_+ denotes the number of +1 links and $W_t = N(N-1)/2$ is the total number of links in the system). Naturally, many different distributions of the propensities are possible for a fixed *q* density. We denote with *r* the relative size of the largest cluster in the optimal clustering, and we consider this parameter as order parameter for the CC problem. Mathematically, this order parameter is defined as

$$r(q) = \left\langle \left\langle \max_{(j)} \left\{ \frac{C_x\{j,q\}}{N} \right\} \right\rangle_{\deg} \right\rangle_x , \quad (3)$$

where $C_x\{k,q\}$ denotes the number of agents in cluster *j,* for an *x* realization of the propensities (distribution of the J_{ij} interactions) with a given *q* density of the positive links. Since the ground-state might be degenerated (different clustering gives the same minimum *K* value), first an average over all these degenerated states are considered, then a second average over the disorder *x* is performed. For finite system sizes, *N,* one can compute the *r(q)* curves using different numerical or analytical approximation techniques. As an example on Figure 1 we present the results of a simulated annealing approach done in [3].

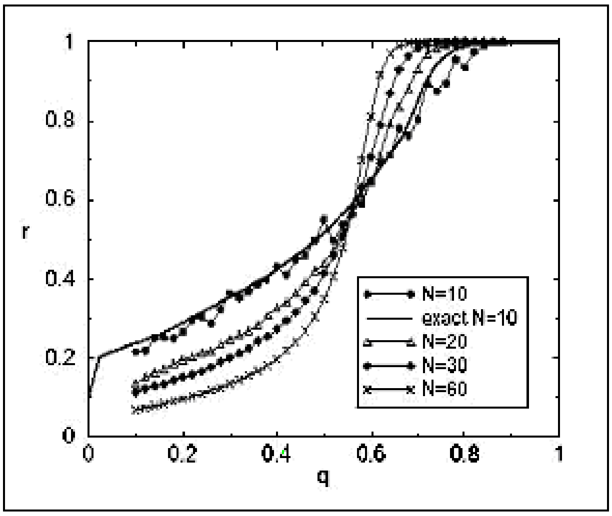

Fig. 1. Simulated annealing results for r(q) in the case of globally coupled system. Results from [3]. The “exact” results are obtained with a total enumeration of all the possible states.

The r(q) curves plotted in Figure 1 suggest a phase-transition like behavior near the $q_c = 0.5$ critical density. This result is in agreement with the predictions discussed in [3] for the thermodynamic limit. For $q < q_c = 0.5$, when there are more negative propensities than positive ones, the order parameter is $r = 0$, suggesting that the relative size of the largest cluster is 0. This corresponds to the case when all elements are in different clusters ($r = 1/N$, which for $N \to \infty$ yields r=0). For $q > q_c = 0.5$ there are more positive links than negative ones. The predicted order parameter is $r = 1$, suggesting that the relative size of the largest cluster is 1. This means that all agents are in the same cluster. The $r(q)$ curves for finite system sizes will have a sharp increase in the neighborhood of the $q_c = 0.5$ critical point. As it is expected for a real phase-transition like phenomena the curves for increasing system sizes are converging to a step-like form in the neighborhood of q_c.

Another clue suggesting that we deal with a phase-transition like phenomenon is from the fluctuation of the order parameter, quantified by the value of its standard deviation:

$$\Delta r(q) = \sqrt{\left\langle r^2(q) \right\rangle_{\deg,x} - \left\langle r(q) \right\rangle^2_{\deg,x}} \ . \tag{4}$$

The average $\langle \ \rangle_{\deg,x}$ is taken for all the performed numerical optimization experiments when an *x* realization of the J_{ij} propensities is given and several *x* realizations (of the order of hundreds) of the links with fixed *q* value. In Figure 2 we present the simulated annealing results for $\Delta r(q)$ taken from [3], considering various system sizes, N. As it is expected for real phase-transitions, a clear maximum is obtained in the vicinity of $q_c = 0.5$, suggesting an increase of the fluctuations at the transition point. The maximum becomes sharper and its position converges towards the $q_c = 0.5$ critical point as the system size is increased.

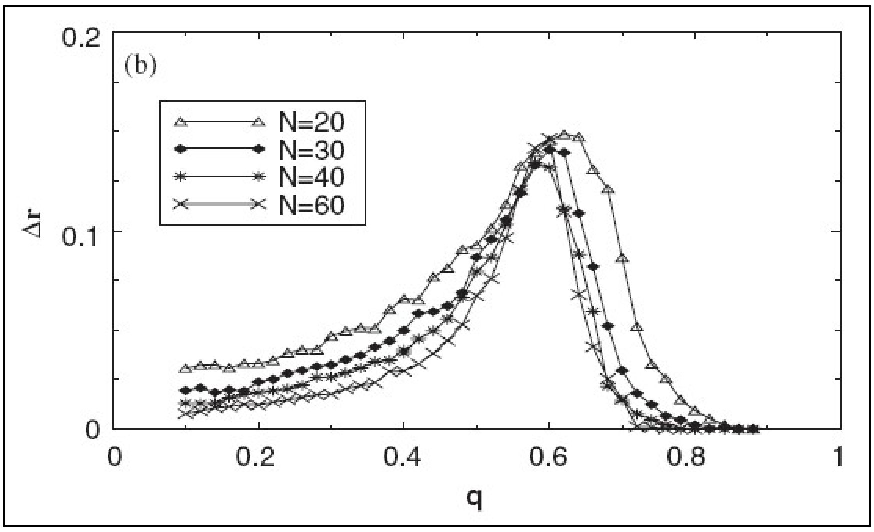

Fig. 2. Standard deviation, $\Delta r(q)$ of the order parameter for different system sizes. Simulated annealing results from [3].

Globally coupled networks (complete graphs) are however rarely relevant to natural and sociological systems [16-18]. Links between the agents in a real natural or sociological system has a quite complex topology, and leads to several non-trivial network structures [18]. A natural question arising then is to consider the CC problem on such random graphs and see how the results discussed up to now will modify. Without entering in details (for a detailed discussion see our recent work [5]) we mention here the basic result for the CC problem on such networks. We will limit the overview only on single component networks in which each node can be reached from each other node by following the existing links. Moreover, we consider the case of randomly diluted Erdos-Renyi type networks with a fixed finite dilution rate (dilution rate is defined here as the number of existing links in the network over the number of possible links). The main results are the following:

1. The *r(q)* curves exhibit a similar trend with the one obtained on complete graphs. There is the same critical proportion of the positive links, $q_c = 0.5$, where the *r(q)* curves have a sharp increase and an inflexion point.

2. The $\Delta r(q)$ curves exhibit a similar trend with the one obtained on complete graphs. At the critical q_c value there is a clear maximum, which increases with the system size and its position is converging for $q_c = 0.5$ as in the case of the globally connected networks.

3. For an infinite size network with a fixed dilution rate ($N \to \infty$), we get the same step-like transition as in the case of globally connected graphs.

In the present study we consider finite graph structures. The results mentioned in items 1.-2. are of importance hereafter.

CORRELATION CLUSTERING AND LOGICAL LEARNING

Any study subject can be represented with a complex graph-like structure. The syllabus can be divided into basic components (parts) which are logically inter-connected in a graph-like manner. Learning logically the subject of study means, that these links are revealed and transformed in positive connections. The positive connection suggests in such case that those parts are logically connected in the students mind. Contrary, a non-logical learning means that the student is assimilating the components without revealing the underlying connections. The link in such cases can be considered negative, since parts of the syllabus are conflicting, and the desired logical relation is not revealed. As an example, in Figure 3a we have considered an idealized study subject composed of 8 parts, interconnected in the depicted graph-like manner. Connections are considered without direction, implication from any node is a link. Depending on the learning style some of the connections are understood and parts of the syllabus becomes connected. If the student uncovers all connections between parts of the material (Figure 3b) the whole syllabus is merged and the learning style is logical. On the other hand if none of the existing connections are revealed the material falls in parts and learning is non-logical (Figure 3c). In general, a situation between these two extremes will occur: some links are revealed (positive in the CC problem) and some links are unexplored (negatives in the CC problem), this is illustrated in Figure 3d. The logical learning rate is a measurable psychological quantity [19], characterizing how logically one assimilates a given material. There are several tests available for quantifying this aspect of learning. Using the graph picture shown on Figure 3 one could interpret the Logical Learning Rate (LLR) as the largest relative size of the syllabus in which the parts are mostly non-conflicting. This is however nothing else but the relative size of the largest cluster in the correlation clustering problem. One can construct thus a useful analogy between the CC problem and logical learning of a material. In the view of this analogy the order parameter, *r*, defined in CC will characterize the LLR of the give subject of study. For continuing the analogy with the CC problem one has to quantify in psychological terms the ratio of positive links, which is the *q* parameter in CC. Our basic hypothesis is that the well quantifiable IQ of students is a good candidate for this. As the intelligence level is higher, existing logical connections between parts of the syllabus are easier to reveal. This means that the ratio of positive links in the graph should depend on IQ and the simplest assumption is: $q \sim IQ$. Since theoretically there is no upper limit for IQ, the proportionality constant is unknown.

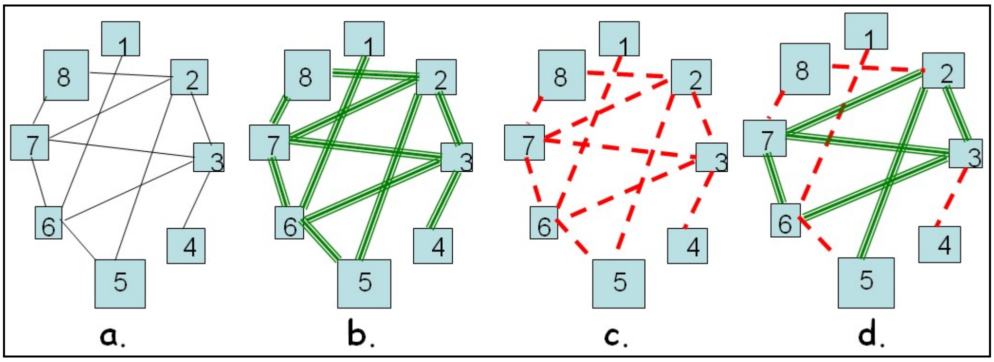

Fig. 3. (a) Graph-like representation for the logically connected parts of a subject of study. (b) When all links are positive (revealed) learning is completely logical and the syllabus is understood as a whole. (c) When all links are negative learning is non-logical and the study material falls in parts. (d) A general situation.

Assuming that the analogy discussed above is justified, one would expect that LLR as a function of IQ should have a similar trend as the *r(q)* curves in the CC problem. This means that the value of LLR should increase sharply at a critical IQ level. Based on conjectures from psychology, this point is expected around an IQ of 120-130, which corresponds to the level of superior intelligence ("gifted" persons) [20]. The variance of LLR as a function of IQ should also exhibit a maximum at this critical IQ value. In the following psychological test measurement results for LLR and IQ for the same control set will be presented and discussed.

MEASUREMENTS AND RESULTS

With the help of a psychologist team we have considered test measurements on a sample of N=177 high-school students. IQ was measured by means of the classical standardized Raven test [21]. The Raven Progressive Matrices (RPM) tests are made up of a series of diagrams or designs with a missing component. Individuals taking the tests are expected to select the correct element to complete the designs from a number of options printed beneath. The instrument has been adapted for Romanian population and has very good psychometric properties. The value of LLR was quantified with the help of the MEMLETICS [22]. The MEMLETICS test is a questionnaire which, based on an evaluation, gives a reasonable estimate of the learning styles relevant for an individual. These learning styles are categorized as: Social, Solitary, Visual, Aural, Verbal, Physical and Logical. We were primarily interested in the results concerning the rate of logical learning (for details see [14]). Results of both tests were summarized and the desired $LLR(IQ)$ and $\Delta LLR(IQ)$ curves were plotted.

First the considered sampling, its correctness and significance level were tested. For this the distribution of the measured IQ values in form of a normalized histogram with bin sizes $d(IQ)=10$ was considered. From the experimental results it was possible to determine the mean IQ value and standard deviation: $\langle IQ \rangle = 112.92; \Delta(IQ) = 14.94$. The normalized histogram was then plotted together with the normal distribution expected with these parameters (Fig. 4):

$$G(y) = W \cdot d \frac{1}{\sqrt{2\pi \cdot \Delta(IQ)}} \exp\left(- \frac{y^2}{2 \cdot \Delta(IQ)} \right) \tag{6}$$

The reasonable agreement of the experimental results with the normal distribution suggests that our sampling is an acceptable one.

The measured LLR values as a function of the IQ values are shown in Figure 5.

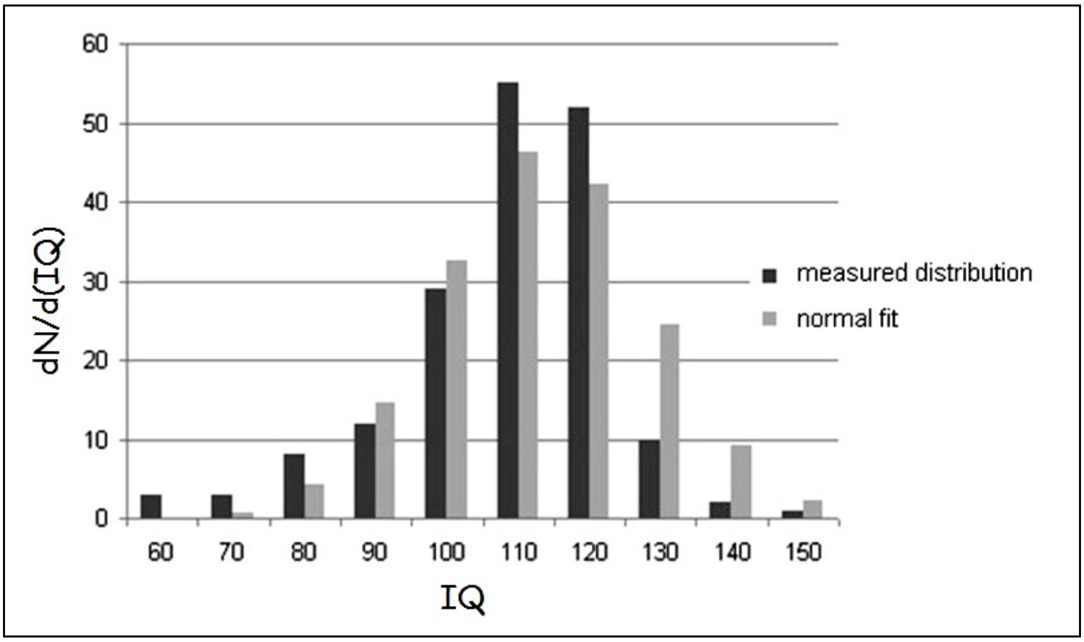

Fig. 4. Distribution in form of a histogram of the measured IQ values. The grey bars indicate the fitted normal distribution.

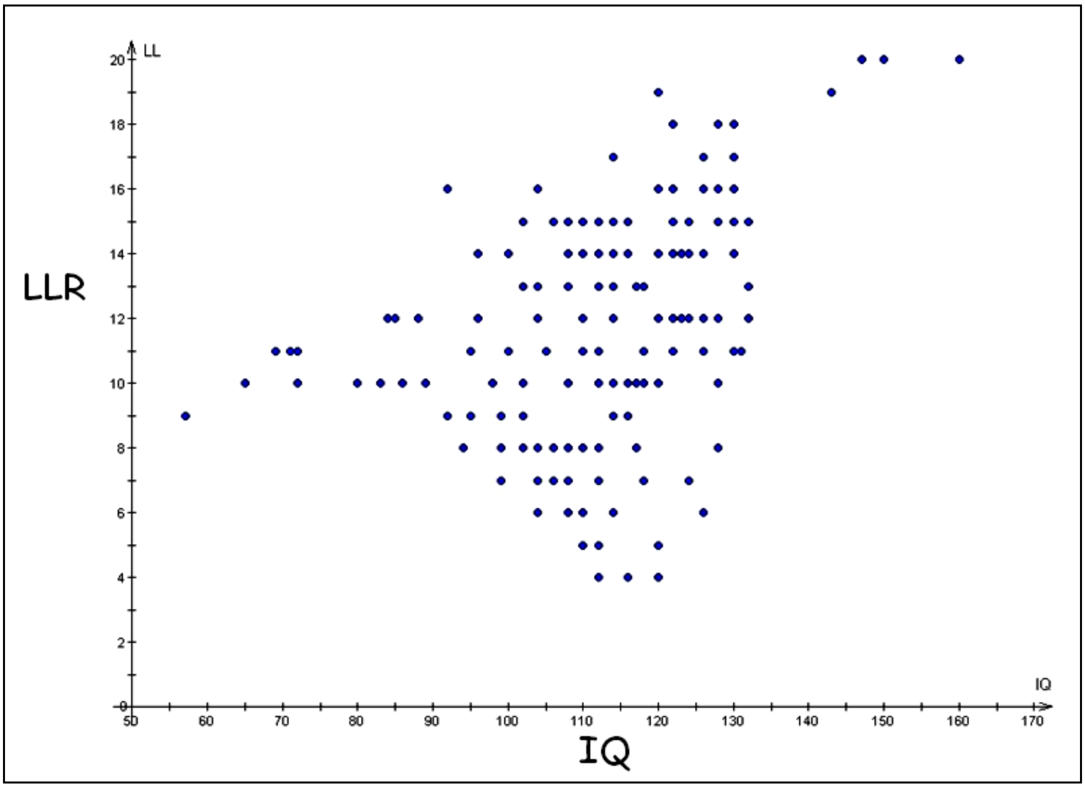

Fig. 5. Plot of experimentally obtained LLR values as a function of IQ.

A simple qualitative view on the set of points suggests that the standard deviation of the LLR values has a maximum somewhere in the $100 \le IQ \le 130$ interval. The dispersion of the experimental points is too high however to get any hint from here on the nature of the $LLR(IQ)$ and $\Delta LLR(IQ)$ curves. An averaging was performed, which resulted in a much smoother and easier interpretable curve. The averaging was done by considering IQ intervals (bins) of size 10 for $IQ < 100$ and $IQ > 130$ and bins of size 5 for $100 \le IQ \le 130$. The non-uniform binning is motivated by the fact that the results are less abundant for $IQ < 100$ and $IQ > 130$ then for $100 \le IQ \le 130$ and so in order to obtain a reasonable average several experimental points are needed. The averaged results are plotted in Figure 6. The trend illustrated by the continuous line resembles well the conjectured shape. In the vicinity of $IQ = 120$ a sharp increase in the LLR level is detectable.

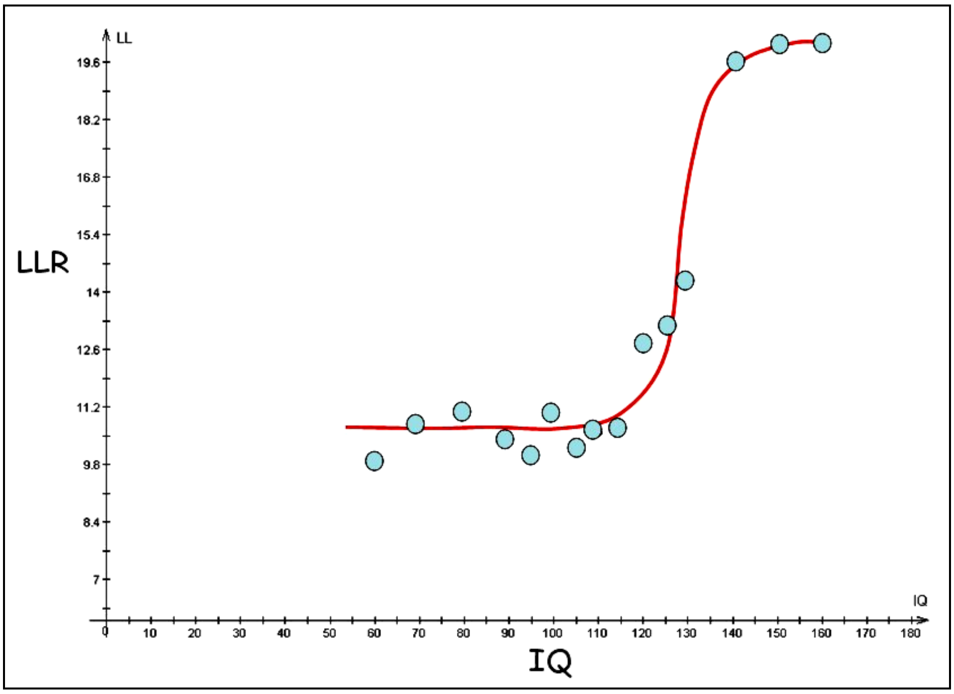

Fig. 6. Bin-averaged results for LLR as a function of IQ. The continuous line illustrates the trend.

For plotting the experimental $\Delta LLR(IQ)$ curve, the standard deviation of the LLR values for individuals with IQ values in the considered bin were calculated and plotted as a function of the average IQ in the bin. Results are given in Figure 7. Similarly with Figure 6, the trend is illustrated by a continuous line. The clear maximum in the vicinity of $IQ = 120$ supports again our conjecture and the applicability of the CC problem for logical learning.

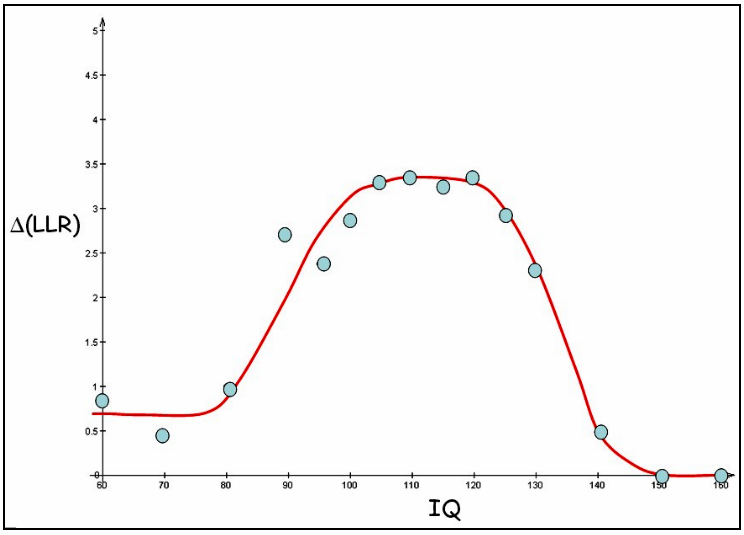

Fig. 7. Averaged experimental results for the standard variation of LLR as a function of IQ. The continuous line illustrates the trend.

CONCLUSIONS

A simple analogy between the Correlation Clustering (CC) problem and logical learning of a subject of study was formulated. The main conclusion which resulted from this analogy is that the Logical Learning Rate (LLR) as a function of the Intelligence Quotient (IQ) should present a sharp increase at a critical IQ level. One would also expect that the standard deviation of LLR as a function of IQ presents a clear maximum at the critical IQ value. Experiments were designed to prove this conjecture. IQ was measured by a standardized Raven test and LLR was quantified with the help of the Memletics test. The obtained results revealed that the conjectured analogy holds for a random sample of individuals and LLR as a function of IQ varies as predicted by CC. From a psychological point of view we replicated previous results concerning connections between logical learning and IQ, by using new measures and models. Additionally, the modeling process, revealed a novel critical transition point, corresponding to the IQ shift from high intelligence to superior intelligence (120-130), with an impact on logical learning performance. This is a new discovery revealed by modeling psychological data in a statistical physics paradigm, with impact on applied psychology (e.g., different teaching styles for people of different IQ categories).

ACKNOWLEDGMENTS

Research supported by a Romanian PN2II Idei 2369/2008 research grant. Financial support for Gy. Cseh was provided by the Sectoral Operational Programme Human Resources Development, Contract POSDRU 6/1.5/S/3 Doctoral Studies: "Through science towards society".

REFERENCES

1. Applications of Statistical Physics, Editors: A. Gadomski, J, Kertész, H.E. Stanley and N. Vandewalle, Proceedings of the NATO Advanced Research Workshop, Technical Universitz of Budapest, 19-22 May, 1999 (North Holland, 2000).
2. N. Bansal, A. Blum and S. Chawla, Machine Learning, vol. 56, pp. 89-113 (2004).
3. Z. Neda, R. Florian, M. Ravasz, A. Libal, and G. Gyorgyi, Physica A, **vol.** 362, **pp.** 357-369 (2006).
4. R. Sumi, and Z. Neda, International Journal of Modern Physics C, **Vol.** 19, **pp.** 1349-1358 (2008).
5. Z. Neda, R. Sumi, M. Ercsey-Ravasz, M. Varga and B. Molnar. Journal of Physics A: Mathematical and Theoretical, vol. 42, pp. 345003-345017 (2009)
6. M.J. Kearns, R.E. Schapire, L.M. Sellie, Machine Learning, vol. 17, pp. 115-141 (1994).
7. R. Axelrod, and S. Bennett, British Journal of Political Science, 23, **pp.** 211-233 (1993).
8. S. Galam, Physica A, **vol.** 230, **pp.** 174-188 (1996).
9. S. Galam, Physica A, vol. 238, **pp.** 66-88 (1997).
10. D. Elderfield, and D. Sherrington, J. Physics C: Solid State Physics, **vol.** 16, **pp.** L497-L503 (1983).
11. A. Erzan, and E.J.S. Lage, J. Physics C: Solid State Physics, **vol.** 16, pp. L555-L560 (1983).
12. T.R. Kirkpatrick and D. Thirumalai, Physical Review B, vol. 37, **pp.** 5342-5350 (1988).
13. P. McGeorge, J.R. Crawford, and S.W. Kelly, Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 23, pp. 239-245 (1997).
14. A.S. Reber, F.F. Walkenfeld and R. Hernstadt, Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 17, pp. 888-896 (1991).
15. M.R. Garey, and D.S. Johnson, Computers and Intractability. A Guide to the Theory of NP-Completeness (New York: W. H. Freeman and Company, 1979).
16. R. Albert and A.L. Barabasi, Rev. Mod. Phys. **Vol.** 74, **pp.** 47-97 (2002).
17. A.L. Barabasi and E. Bonabeau, E., Scientific American, **vol.** 288, 60-69 (2003).
18. A.L. Barabasi et. al, Physica A, **vol.** 311, **pp.** 590-614 (2002).
19. J. Anderson**,** Cognitive psychology and its implications (6th ed.) (New York: Worth Publisher, 2004).
20. D. Wechsler, The Measurement of Adult Intelligence. Baltimore (The Williams & Wilkins Company, 1944)
21. J. Raven, Cognitive Psychology, vol. 41, pp. 1-48 (2000).
22. MEMLETICS test: http://www.memletics.com/