Final Report on EPSRC Research Grant GR/N00562
Irregular Gallager-type error-correcting codes -
a statistical mechanics perspective

David Saad
Aston University, Birmingham B4 7ET

1 Background

Error-correcting codes are of significant practical importance as they provide mechanisms for retrieving the original message after corruption due to noise during transmission. They are being used extensively in most means of information transmission from satellite communication to the storage of information on hardware devices. The coding efficiency, measured in the percentage of informative transmitted bits, plays a crucial role in determining the speed of communication channels and the effective storage space on hard-disks. Rigorous bounds [1] have been derived for the maximal channel capacity for which codes, capable of achieving arbitrarily small error probability, can be found. However, until recently, the performances of most existing practical error-correcting codes have been significantly below this bound; and although in the last couple of years the gap is gradually closing, the search for more efficient error-correcting codes is still on.

Low Density Parity Check (LDPC) codes, is one family of error correcting codes that enjoyed a significant level of popularity in recent years. It was originally introduced by Gallager [2] and re-introduced by MacKay and Neal [3], showing excellent performance with respect to most existing codes. In fact, it has recently been shown that irregular constructions of LDPC codes provide better performance than any other method [4, 5] including the celebrated turbo codes [6]. One of the most significant advantages of LDPC codes over turbo codes is that they are amenable to analysis, and their performance can therefore be systematically improved. LDPC codes are generally based on the introduction of random sparse matrices for generating the code word as well as for decoding the corrupted code-word. Various decoding methods have been successfully employed; they mainly rely on methods adopted from graphical models such as belief propagation [7, 8] and belief revision [9].

In previous studies we have used methods of statistical physics, devised in the field of diluted spin systems, to investigate the performance and efficiency of LDPC error-correcting codes [10, 11]. The analytical results provide insight into the encoding/decoding mechanism and its dependence on the choice of parameters. This insight was further exploited for suggesting specific constructions [5] of very high performance. As the decoding method used plays a critical role in defining the method's efficiency, we also studied the most commonly used decoding method for state of the art error-correcting codes, belief propagation. We discovered the similarities between belief propagation techniques and the Bethe approximation [12] of statistical physics.

The current project relied mainly on research carried out by the principal investigator and the post-doctoral researchers supported by the grant, but also benefited from work carried out by collaborators and visitors.

Most of the original project objectives have been achieved as will be described below. Objective (4), aimed at devising improved decoding/encoding methods, has been modified after the introduction of a new decoding methods [13] by another group and the realisation that it will be very difficult to provide a superior *practical* decoding method. Instead, we extended the study of LDPC codes to areas of multi-user information theory communication problems.

The main achievement of this study is in continuing the successful application of statistical physics methods to the study of LDPC codes. This analysis provides new insight and results that

cannot be obtained using methods employed in the information theory community, especially in the case of finite connectivity codes. More specifically we studied both practical and theoretical properties of general LDPC codes, regular and irregular (objectives 1,2); investigated the decoding process and its relation to other methods used in the fields of statistical physics and graphical models (objective 3); showed how specific irregular LDPC codes can be optimised (objective 2); and extended results obtained for infinitely large codes to the case of large but finite systems by studying properties of the *reliability exponent* in LDPC codes, which provides the block error probability in large systems; in addition, we studied the typical performance of LDPC codes in multi-user communication scenarios such as CDMA (objective 4), and carried out research in a new, but theoretically related, direction of computational complexity (graph colouring).

2 Key Advances and Supporting Methodology

The study is based on the statistical physics formulation of LDPC error correcting codes. In a general scenario, a message represented by an N dimensional binary vector $\boldsymbol{s} \in [0,1]^N$ is encoded, using a generator matrix $\boldsymbol{G}$, to an M dimensional binary vector $\boldsymbol{t} = \boldsymbol{G}^T \boldsymbol{s}$ which is then transmitted through a noisy channel. The noise corrupting the transmission can take different forms; in this study we mainly considered the Binary Symmetric Channel (BSC), Additive White Gaussian Noise (AWGN), Laplace channel, Binary Erasure Channel (BEC) and general symmetric channels. For explaining the framework we will consider first a binary symmetric channel with some flip probability p per bit, represented by a binary vector $\boldsymbol{n} \in [0,1]^M$ with a unit entry whenever a flip corruption occurs. The received message $\boldsymbol{r} = \boldsymbol{t} + \boldsymbol{n}$ (mod 2) is then decoded to retrieve the original message using another matrix $\boldsymbol{H}$.

One can identify several slightly different versions of LDPC codes. The two examined in this project are that of Gallager and MN codes [3]. Both are based on the choice of two randomly-selected sparse matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of dimensionality $(M-N)\times N$ and $(M-N)\times(M-N)$ (Gallager), and $M\times N$ and $M\times M$ (MN) respectively. These matrices are characterized by the fixed number of non-zero unit elements per row and per column in regular codes and their distribution in irregular codes.

The difference between the two codes is related to the generator and decoding matrices. A Gallager code is defined by a binary matrix $\boldsymbol{H} = [\boldsymbol{A} \mid \boldsymbol{B}]$, concatenating two sparse matrices known to both sender and receiver; while the generator matrix takes the form $\boldsymbol{G} = [\boldsymbol{I} \mid \boldsymbol{B}^{-1}\boldsymbol{A}]$ (mod 2), $\boldsymbol{I}$ being the identity matrix. The received corrupted vector is decoded by obtaining the syndrome vector $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{r}$ which reduces the decoding problem to finding the most probable solutions to the equation $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{\tau}$ (mod 2) , where $\boldsymbol{\tau}$ represents the noise vector. Finding a good estimate of the vector $\boldsymbol{n}$ enables one to retrieve the original message $\boldsymbol{s}$. This estimate is typically being obtained using methods developed in the area of graphical models, which are somewhat similar to methods used in statistical physics (the Bethe approximation and/or the TAP approach). The corresponding decoding matrix in MN codes is the matrix $\boldsymbol{B}$ itself, providing a similar equation of the form $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{\sigma} + \boldsymbol{B}\boldsymbol{\tau}$ (mod 2) , where $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ represent the signal and noise vector variables respectively.

The theoretical performance of a code is characterised by the maximal noise level below which corrupted messages can, in principle, be perfectly retrieved; this is always bounded from above by Shannon's theoretical limit which does not provide any information as to how this performance can be achieved. The practical performance of a code is defined by the maximal noise level below which corrupted messages can be perfectly retrieved in practical time scales and computing resources. In previous studies [14] we identified the theoretical critical noise level with the thermodynamic transition where the ferromagnetic state ceases to be dominant; and the practical limiting noise level with the dynamical transition, where the ferromagnetic state ceases to be the only solution.

The performance of codes (both practical and theoretical) depends on the code construction, i.e., the number of non-zero unit elements per row/column in the various matrices. In regular Gallager code we consider the number of unit elements per row/column in the matrix $\boldsymbol{H}$ while in MN code we consider separately the number of non-zero (unit) elements per row/column in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. In irregular constructions we investigate the *distribution* of these connectivity values.

The theoretical framework (for the BSC) has been developed for various LDPC code constructions in several separate papers [10, 11, 14, 15]; a comprehensive and coherent presentation of the theoretical framework for LDPC codes, the methods used and further progress in the analysis of irregular constructions (e.g., that of [5]) has recently been published as a book chapter [16](O1). The basic framework has also been studied in [17].

Extending the statistical mechanical framework to the case of real channels is more challenging due to the real valued nature of the noise and received codeword vectors. We used a technique developed in the information theory community (e.g., in [3]), which effectively maps the decoding problem back onto a binary channel; we employed this method to study the AWGN and Laplace channels [18]. The results obtained in our study show qualitatively similar behaviours to those obtained for BSC in both Gallager and MN codes (i.e., the type of solutions and transitions found), while details of the observed performance vary from channel to channel. Some generic results have been obtained for *general* symmetric channels [18]. A research activity that focused on the BEC, by Dr. Malzahn, was terminated after a similar research by another group was made public [19].

The basic theoretical framework has also been extended to non-binary alphabets; for instance, to the case of LDPC codes over Galois fields (O1). Irregular codes in this representation, which is commonly used within the coding community, revealed a very high performance [20], what triggered our interest in non-binary alphabets. We studied the effect of non-binary alphabets of this type on both theoretical and practical critical noise levels in regular LDPC codes [21]. The results show that codes of this type saturate Shannon's limit as their connectivity value become infinite, irrespective of the alphabet used; while for finite column connectivity their behaviour critically depends on the connectivity value chosen. For column connectivity of 3 and above, the theoretical limit is monotonically improving as the alphabet becomes more complex while the practical decoding limit deteriorates. Codes of column connectivity 2 exhibit a continuous transition from optimal to sub-optimal solutions at a certain noise level, below which practical decoding dynamics converges to the optimal, while their optimal decoding performance is generally inferior. These results are in agreement with numerical results reported in the literature.

A different approach to the study of both practical and theoretical properties of LDPC codes, using a microcanonical ensemble, has been carried out in [22, 23, 24]. In this approach, which shows some similarity to methods employed in the information theory literature, one is looking at the number of solution vectors, given the parity-check constraints; these, in turn, are defined by the choice of matrices $\boldsymbol{H}$ (in Gallager) or $\boldsymbol{A}$ and $\boldsymbol{B}$ (in MN). The number of solutions obtained (entropy) under given shells of overlap with the true noise (or/and signal) vectors (weight enumerator) or magnetization values (magnetization enumerator), are then used for finding the critical noise level values obtained by several decoding schemes: Maximum A Posteriori (MAP), typical set decoding and Marginal Posterior Maximizer (MPM). The results obtained, provide insight into the relations between the various decoding methods (O3), and show the same critical theoretical noise level values for all three methods.

The critical noise values obtained also agree with results we obtained by different methods in [16] and [25, 26]. A study dedicated to typical set decoding in LDPC codes has also been reported in [27] (O3). In this decoding scheme, error occurs either when the transmission is corrupted by an untypical noise or when two or more typical sequences satisfy the parity check equation. We showed that the average error rate for the latter case over a given code ensemble can be tightly evaluated

using methods of statistical physics, as well as its dependence on the message length. The results obtained show more optimistic values than the bounds derived in the information theory literature.

Having identified the practical limiting noise level with the dynamical transition, characterised by the emergence of new sub-optimal solutions in addition to the thermodynamically dominant ferromagnetic state; one can then optimise the code construction by examining the corresponding spinodal point, marking the emergence of new solutions. Using this method we optimised a specific irregular construction, suggested in [5], which includes submatrices with two different row connectivity values; the construction was optimised with respect to a single parameter which marks the balance between two submatrices (having different row connectivity values) [16]. The optimal irregular code we obtained is in full agreement with the optimal irregular codes obtained numerically (O2). Extending the method beyond a few parameters, i.e., to full distributions, remains difficult, and is likely to be restricted to the results obtained by density evolution [4].

Most of the analysis carried out so far focused on the infinite message length, although most practical messages and codewords are long, but finite. Finite code induce a certain block error probability below the theoretical limits even if exhaustive decoding methods are being used. The error probability decays exponentially with the system size; the exponent, which depends on the code's properties is termed the *reliability exponent*. To study the performance of finite codes, depending on the selected code structure, we calculated the *typical* reliability exponent for Gallager codes [25, 26] using methods of statistical physics. The results obtained are in full agreement with those obtained in the information theory literature for infinite row/column connectivity codes, but also offers typical reliability exponent values for *finite* connectivity codes, which cannot be obtained using methods used in the information theory community (O1).

Another objective of the research programme focused on the relation between graphical decoding methods such as belief propagation [7, 8] and belief revision [9], and methods developed in the statistical physics literature. We investigated this relation in a number of manuscripts [16, 28, 29], finding high similarity between the Bethe approximation of statistical physics and graphical decoding methods (O3). This relation has since been well established. Our plan to improve existing decoding methods by accounting for loops, which are not considered in the current approximation, has been cancelled once we learned of the survey propagation approximation that has only recently been presented [13].

We therefore modified objective (4), aimed originally at devising improved decoding/encoding techniques; instead we concentrated on the study of LDPC and decoding methods in areas of multi-user communication; in particular as part of a CDMA system. This research focused initially on the use of density-evolution, the information theory equivalent of the Bethe approximation, in a multi-stage multi-user detection/decoding as part of a CDMA scheme [32, 33], showing significant improvement over conventional decoding techniques [31]. The analysis of the suggested method has been carried out using methods of statistical mechanics.

We then integrated the LDPC and CDMA framework for improved code modulation in two separate configurations (O4). The first [30] focuses on an LDPC based coded modulation scheme, deriving analytical results for arbitrary signal point mapping; numerical results have then been obtained for Gray and set-partitioning mappings, showing that Gray mapping exhibits smaller threshold signal to noise ratio, whereas the set partitioning mapping exhibits close to optimal theoretical upper limit. In the second method we analysed the theoretical/practical performance of coded CDMA systems in which regular LDPC codes are used for channel coding [34]. Also here, we identified both practical and theoretical limitations of the system, showing that one cannot achieve any coding gain in the (impractical) limit of dense matrices; we also point to possible practical configurations where the new scheme is likely to improve the performance.

In addition to the planned part of the programme there was also one project that emerged

spontaneously within the course of research, resulting in a journal publication. This research activity, in the area of computational complexity is based on similar theoretical tools to those used in the study of LDPC codes. The NP-complete problem studied is that of graph coloring [35], one of the basic problems in the family of hard computational problems.

3 Project Plan Review

The project generally progressed according to plan. However, the frequent staff changes and the highly competitive nature of this research area prevented us from achieving all the original goals set at the beginning of the project. Dr. van Mourik, that was originally hired to carry out the research, was offered a lectureship position and left after 11 months; in order to minimise the learning curve and to make the most of the remaining funds, we invited Dr. Tanaka of the Tokyo Metropolitan University to continue the research effort as a visiting research fellow (after getting EPSRC's approval). Unfortunately, Dr. Tanaka could only stay for 8 months due to his teaching duties back in Japan. With the remaining funds we hired Dr. Malzahn for 2 months to carry out a specific and limited task (study properties of LDPC codes in a BEC). In addition, difficulties in optimising the code connectivity distributions restricted the optimisation process to a few parameters; and the presentation of survey propagation [13] terminated our attempts to devise a new decoding method that accounts for construction loops. This was also the reason for modifying objective (4) of the original plan. In addition, studies of the BEC channel carried out by Dr. Malzahn have been stopped after a similar study, by another group, became public [19]. Like in many other research programmes in a very active area, the programme had to be constantly adapted to the changing conditions.

4 Research Impact and Benefits to Society

Being a theoretical project, the *direct* impact and benefit to society is small. The main contribution is in obtaining a better understanding of LDPC codes and their limitations, and in our ability to exploit this knowledge to improve code performance. LDPC codes have only recently started to get into commercial products and their use in conjunction with CDMA systems is in very preliminary stages. It will take time for the society to benefit from this theoretical research.

5 Explanation of Expenditure

The availability of very high performance PCs made the original plan of purchasing a Compaq DS20 6/500 workstation uneconomical. Instead, we purchased a cluster based on fast PCs.

6 Further Research and Dissemination Activities

Results obtained in this project have been published in prestigious journals in this field (such as Physical Review E, Europhysics Letters) and have been presented in the top international conferences (e.g., ISIT2002). Most of the resulting publications, as well as this report may be obtained from our web site `http://www.ncrg.aston.ac.uk/` .

This project gave rise to new questions and research topics. Some of those, which represent a direct continuation to the work carried out in this project, are currently investigated in two collaborative efforts:

• As part of the UK-Japan Joint Project Grant on *Statistical Physics of Disordered and Complex Systems*, funded by the Royal Society (14245), £12,000.

• Part of the European net for *Statistical Physics of Information Processing and Combinatorial Optimization* STIPCO, European Commission RTN2-2001-00197, 116,000 EUR.

References

[1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).

[2] R.G. Gallager, IRE Transactions on Information Theory, **IT-8,** 21 (1962).

[3] D.J.C. MacKay, IEEE Transactions on Information Theory, **45** 399 (1999).

[4] J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, IEEE Trans. on Information Theory **47**, 619 (2001).

[5] I. Kanter and D. Saad, Phys. Rev. Lett. **83**, 2660 (1999).

[6] C. Berrou, A. Glavieux and P. Thitimajshima proceedings of the 1993 IEEE International Conference on Communications, Geneva Switzerland 1064 (1993).

[7] J.Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann), (1988).

[8] B.J. Frey, *Graphical Models for Machine Learning and Digital Communication* (MIT Press), (1998).

[9] Y. Weiss, Neural Computation, **12**, 1 (2000).

[10] Y. Kabashima, T. Murayama and D. Saad, Phys. Rev. Lett. **84**, 1355 (2000).

[11] T. Murayama, Y. Kabashima, D. Saad, and R. Vicente, Phys. Rev. E **62**, 1577 (2000).

[12] Y. Kabashima and D. Saad, Europhys. Lett., **44**, 668 (1998).

[13] M. Mézard and R. Zecchina, `cond-mat/0207194`, submitted (2002).

[14] R. Vicente, D. Saad and Y. Kabashima, Jour. Phys. A **33**, 6527 (2000).

[15] R. Vicente, D. Saad and Y. Kabashima, Europhys. Lett. **51** 698 (2000)

[16] R. Vicente, D. Saad and Y. Kabashima, in *Advances in Imaging and Electron Physics*, 232, Ed. P. Hawkes, Elsevier Oxford (2002).

[17] D. Saad, Y. Kabashima, T. Murayama and R. Vicente, in Cryptography and Coding, 307, Ed. B. Honary, Springer-Verlag, Berlin 2001.

[18] T. Tanaka and D. Saad, *Typical performance of low-density parity-check codes over general symmetric channels*, submitted (2002).

[19] S. Franz, M. Leone, A. Montanari and F. Ricci-Tersenghi, `cond-mat/0205051`, submitted (2002).

[20] D.J.C. MacKay, S.T. Wilson and M.C. Davey, IEEE Trans. on Communications, **47**, 1449 (1999).

[21] K. Nakamura,Y. Kabashima, and D. Saad, Europhys. Lett., **56**, 610 (2001).

[22] J. van Mourik, D. Saad and Y. Kabashima, Phys. Rev. E **66**, 026705 (2002).

[23] J. van Mourik, D. Saad and Y. Kabashima, in Cryptography and Coding, 148, Ed. B. Honary, Springer-Verlag, Berlin (2001).

[24] J. van Mourik, D. Saad and Y. Kabashima, Proceedings of ISIT 2002, 256 (2002).

[25] Y. Kabashima, N. Sazuka, K. Nakamura and D. Saad, Phys. Rev. E **64** , 046113 (2001).

[26] Y. Kabashima, N. Sazuka, K. Nakamura and D. Saad, Proceedings of ISIT 2002, 255 (2002).

[27] Y. Kabashima, K. Nakamura and J. van Mourik, Phys. Rev. E **66** , 036125 (2002).

[28] Y. Kabashima and D. Saad, in *Advanced Mean Field Methods - Theory and Practice*, Eds. M. Opper and D. Saad, MIT press, Cambridge US, 51 (2001).

[29] D. Saad, Y. Kabashima and R. Vicente, in *Advanced Mean Field Methods - Theory and Practice*, Eds. M. Opper and D. Saad, MIT press, Cambridge US, 67 (2001).

[30] T. Tanaka, *Coded-modulation with Gallager code*, submitted (2002).

[31] T. Tanaka, in Proceedings of the workshop on concepts in information theory, 94 (2002).

[32] T. Tanaka, Proceedings of ISIT 2002, 23 (2002).

[33] R. Mueller, G. Caire, and T. Tanaka, Proceedings of the 40th Annual Allerton Conf. on Communication, Control, and Computing, (2002).

[34] T. Tanaka and D. Saad, *A statistical-mechanical analysis of coded CDMA with regular LDPC codes*, submitted (2002).

[35] J. van Mourik and D. Saad, Phys. Rev. E **66**, 056120, (2002).

Preprints may be obtained from the NCRG database `http://www.ncrg.aston.ac.uk/`