Introduction to the Special Issue on Codes on Graphs and Iterative Algorithms

IN the 50 years since Shannon determined the capacity of ergodic channels, the construction of capacity-approaching coding schemes has been the supreme goal of coding research. Finally today, we know of practical codes and decoding algorithms that can closely approach the channel capacity of some classical memoryless channels. It is a remarkable fact motivating this special issue that all known practical, capacity-approaching coding schemes are now understood to be codes defined on graphs, together with the associated iterative decoding algorithms.

Considering the intensity of the efforts to achieve Shannon's limit in the past few decades, it is ironic that the main ideas pertaining to codes on graphs and to sum-product decoding were, in essence, invented 40 years ago by Gallager but were subsequently neglected by the coding community. It is some consolation to recognize that Gallager's codes were ahead of their time—given the limited processing capabilities of the time, Gallager's codes were simply considered impractical.

Even so, a small number of researchers continued to study codes on graphs. Zyablov and Pinsker (1975) and Margulis (1982) studied Gallager's low-density parity-check codes. In 1981, Tanner wrote a landmark paper that formally introduced the graphical model notation for describing codes, proved the optimality of the sum-product algorithm in cycle-free graphs, and founded the topic of algebraic methods for constructing graphs suitable for sum-product decoding.

Recent excitement about codes on graphs and sum-product decoding was ignited in the mid-1990s by the excellent performance exhibited by the turbo codes of Berrou and Glavieux, MacKay and Neal's near-capacity results on Gallager codes, and the linear-complexity expander graph codes of Sipser and Spielman. At the same time, other researchers recognized a unifying theme in the iterative decoding algorithms and in the mid-1990s papers showing the connections between iterative decoding of codes on graphs and algorithms in the artificial intelligence and systems theory communities were published by Wiberg, Loeliger, and Koetter (1995), Frey and Kschischang (1996), McEliece (1997), McEliece, MacKay, and Cheng (1998), Kschischang and Frey (1998), and Aji and McEliece (1998).

By now there exists a number of classes of codes on graphs that can approach the Shannon limit quite closely with moderate complexity, or extremely closely with high but not infeasible complexity. Indeed, ideas presented in this special issue have allowed us to approach the Shannon limit to within hundredths of a decibel.

Publisher Item Identifier S 0018-9448(01)00715-5.

All of these codes consist of multiple, easily decodable, relatively simple component codes connected by a large pseudo-random permutation, usually represented by a bipartite graph. The permutation allows the construction of long codes, but does not of itself require any decoding computations. The decoding consists only of multiple iterative decodings of the simple component codes.

Although some of the questions that arise in the context of codes on graphs are answered in this special issue, many remain open. Most results to date have been experimental—will useful supporting theory emerge? Work in the 1990s focused on simple channels—can this field be extended successfully to include more complicated, realistic channel models? Most analyses so far assumed the code graphs to be effectively cycle-free—can we analyze the more practical case of graphs with cycles? Realizations of codes in the existing literature century focused on random graphs—can we use algebra to construct graphs that have useful properties and at the same time are suitable for iterative decoding? Experiments in the 1990s came from digital simulations—are there more effective implementations of iterative decoding?

Two main problems motivate all papers in this special issue: the representation of systems on graphs and the sum-product algorithm as a means of approximate inference. Some contributions focus on codes and applications of coding theory while others take a more fundamental approach and consider general iterative, graph-based algorithms. Most assume the graph is effectively cycle-free, but some begin to offer theoretical insight into graphs with cycles.

The issue begins with a tutorial paper by Kschischang, Frey, and Loeliger, who compare and contrast a variety of graphical representations and algorithms. The authors observe that all these graphical representations express the factorization of some global function into a product of local functions, and that the associated iterative algorithms are efficient methods of computing a marginal of the global function by iterative computation of local functions. They show that a large variety of known algorithms, including generalizations of the Kalman filter on Gaussian graphs, can be derived as special cases of this general framework.

Geman and Kochanek build on the graphical model framework, and define code structure using Chomsky context-free grammars and production systems. They also show how "coarse-to-fine" dynamic programming and "thinning" can be used to reduce the complexity of exact decoding.

The first paper by Luby, Mitzenmacher, Shokrollahi, and Spielman treats the application of graph-based codes to the erasure channel. The authors show that it is possible to closely approach the capacity of an erasure channel with a simple iterative procedure. One of the key results of this paper is that

this is possible with a time complexity that is linear in the block length of the code.

Two central questions for the performance of iterative algorithms are the existence of fixed points and the conditions under which the algorithms will converge to them. For sum-product decoding on memoryless channels, much progress has been made recently using an approach called *density evolution*, which interestingly also goes back to Gallager. Some of the key papers on this topic appear in this special issue (Luby, Mitzenmacher, Shokrollahi, and Spielman; Richardson and Urbanke; Richardson, Urbanke, and Shokrollahi). These authors prove concentration theorems which show that in the limit large random graphs can be assumed to be effectively cycle-free. This allows the calculation of precise convergence thresholds. Code parameters can then be carefully chosen to optimize these thresholds. This approach has been used to find low-density parity-check codes with performance extremely close to the Shannon limit on typical memoryless channels.

In another contribution, Richardson and Urbanke show how irregular low-density parity-check codes can be constructed so that encoding takes linear time.

The density evolution algorithm can be sped up by approximating the densities of messages by the normal distribution (Chung, Urbanke, and Richardson; El Gamal and Hammons). The iterations of the decoding algorithm can then be modeled as a simple one-parameter dynamical system. Quite accurate and fast performance predictions can be made with this technique.

Davey and MacKay apply powerful low-density parity-check codes to construct schemes for communicating over channels with synchronization errors. Their schemes are capable of correcting hundreds of insertions and deletions per block with reasonably high efficiency, well beyond any such scheme proposed previously.

The theoretical analysis of iterative algorithms in graphs with cycles is pursued in a number of contributions (Agrawal and Vardy; Freeman and Weiss; Frey, Koetter, and Vardy; Rusmevichientong and Van Roy; Sella and Be'ery).

Agrawal and Vardy analyze the dynamics of iterative decoding as a high-dimensional nonlinear system. Despite the enormous dimensionality, it is possible to make some qualitative statements, as well as to outline general analysis methods taken from the theory of nonlinear dynamical systems.

Sella and Be'ery examine the dynamics near the fixed points of iteratively decoding of product codes. Using a geometric framework developed by Richardson, they linearize the turbo decoding dynamics near a fixed point to study issues of convergence and stability. The authors prove that for any 2×2 (information bits) product code, there is a unique and stable fixed point, and give sufficient conditions for stability in the general case.

Freeman and Weiss show that the fixed points of the max-product variant of the sum-product algorithm (which reduces to the Viterbi algorithm in a trellis) are local maxima according to a particular neighborhood function. Rusmevichientong and Van Roy show that when the sum-product algorithm is applied in a graph describing the product of two Gaussian kernels, the fixed points give the correct means. These are two of the few general results currently known for the sum-product algorithm in graphs with cycles.

Frey, Koetter, and Vardy study the behavior of iterative decoding algorithms as geometric algorithms in signal space. They show how the signal space can be populated with points that correspond to different decoding decisions and they give formulas for computing these points, whose number grows doubly-exponentially with the number of decoding iterations. They provide visualizations of low-dimensional codes to illustrate the geometry of iterative decoding in signal space.

Another major thrust of this special issue is to describe and characterize the properties of codes on graphs. It is by now well understood that the expansion property is of central importance for iterative decoding. Burshtein and Miller consider the behavior of message passing algorithms once they have corrected most errors. They show that if the graph of the low-density parity-check code is a sufficiently good expander, then these algorithms will definitely converge once they have corrected a sufficiently large number of errors.

Zémor simplifies and improves the hard-decision decoding algorithm used in the expander graph codes of Sipser and Spielman. The decoding algorithm generalizes the natural iterative hard-decision decoding algorithm for product codes.

Random graphs are known to be good expander graphs with high probability, which partly explains the success of random graphs in the construction of iteratively decodable codes. However, there is high practical interest in code constructions that optimize the tradeoff of performance versus complexity. Some papers in this special issue present new classes of codes with good performance–complexity tradeoffs. Ping and Wu develop a new class of low-complexity codes on graphs with performance comparable to turbo codes, but with significantly less decoding complexity. Ping, Huang, and Phamdo present analytical and simulation results on a class of codes based on so-called "zigzag" graphs. This class is both simple and powerful and is similar in spirit to the recent work of Divsalar, Jin, and McEliece on "Repeat–Accumulate" codes.

Pseudorandom constructions of codes on graphs work well, particularly for long block lengths, but more structured algebraic constructions are desirable both in order to describe codes compactly and to control their distance and graph-theoretic parameters more precisely.

Tanner shows how to derive minimum-distance bounds for codes on graphs. His approach is quite general, and is computationally feasible for structured graphs that are not too large.

The paper by Banihashemi and Kschischang extends the fundamental algebraic theory of Tanner graph representations to Abelian group codes and lattices, just as the theory of trellises for linear codes was earlier extended to group codes and lattices.

Forney introduces "normal realizations" for systems, which can be derived from a graph of Wiberg type by requiring that all variables have degree at most two. Any state realization of a code can be put into normal form without essential change in the corresponding graph or in its decoding complexity. Forney shows that the sum-product algorithm can be applied to normal realizations, and gives a proof of the Cut-Set Bound, which shows that graphs with cycles may represent a given system with much smaller local complexity than a cycle-free representation.

Forney also develops striking duality results for normal group realizations, proving that a local dualization procedure applied to a normal realization yields a normal realization for the dual.

The distributed graphical structure of the sum-product algorithm maps naturally to parallel very large scale integration (VLSI) implementations. Loeliger, Lustenberger, Helfenstein, and Tarköy show that the local computations performed in the sum-product algorithm map naturally to analog transistor circuits. The potential speed–power improvements of the resulting analog VLSI implementations are enormous.

The sum-product algorithm is a general inference algorithm and it is potentially very useful for inference in joint estimation and detection models. Worthen and Stark review how graphical models can be used in the design of iterative receivers, and introduce some common principles for combining functions. The editors believe that such a joint approach to various subtasks in a complex system will be important for the development of sophisticated future communications systems.

As Guest Editors for this special issue, we wish to thank all those who submitted manuscripts to the special issue. We also thank our many thoughtful, diligent, and timely referees. We received more good papers than could be accommodated in this special issue, and have given preference to those that focus on fundamental theoretical problems or promising new directions. We begin the 21st century with this special issue on codes on graphs and iterative algorithms, which we hope will lay significant stepping stones toward answering some major existing questions and opening up new directions of research.

BRENDAN J. FREY, *Guest Co-Editor in Chief*

RALF KOETTER, *Guest Co-Editor in Chief*

G. DAVID FORNEY, Jr., *Guest Editor*

FRANK R. KSCHISCHANG, *Guest Editor*

ROBERT J. MCELIECE, *Guest Editor*

DANIEL A. SPIELMAN, *Guest Editor*

Brendan J. Frey (M'00) was born on August 29, 1968, in Calgary, AB, Canada. He received the B.Sc. degree in electrical engineering from the University of Calgary, in 1990, the M.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1993, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1997

He worked as Research Scientist at Bell-Northern Research in Ottawa, ON, Canada, from 1990 to 1991. From 1997 to 1999, he was a Beckman Fellow at the University of Illinois at Urbana-Champaign, where he continues to be an Adjunct Faculty Member in Electrical and Computer Engineering. He is now an Assistant Professor in the Department of Computer Science at the University of Waterloo, Waterloo, ON, Canada. In June 2001, he will become the Director of the new Intelligent Algorithms Laboratory at the University of Toronto. He has given over 30 invited talks, and published over 60 papers on inference and estimation in complex probability models for signal processing, iterative decoding, machine learning, computer vision, automated diagnosis, data compression, and pattern recognition. In 1998, MIT Press published his book, *Graphical Models for Machine Learning and Digital Communication*. In addition to leading his group at the University of Waterloo, he supervises graduate students and co-leads projects at the University of Illinois at Urbana-Champaign, and consults on a regular basis at Microsoft Research, Redmond, WA.

Dr. Frey was awarded the Ontario Premier's Research Excellence Award "for inventing new inference algorithms and applying them to practical problems in computer vision, signal processing and digital communications" in September 2000.

Ralf Koetter (S'92–M'95) was born in Königstein, Germany, on October 10, 1963. He received the Diploma in electrical engineering from the Technical University Darmstadt, Germany in 1990. He received the Ph.D. degree for his work in the field of algebraic-geometric codes at the Department of Electrical Engineering at Linköping University, Linköping, Sweden.

From 1996 to 1997, he was a Visiting Scientist at the IBM Almaden Research Laboratory, San Jose, CA. He was a Visiting Assistant Professor at the University of Illinois at Urbana-Champaign and Visiting Scientist at CNRS in Sophia Antipolis, France during the years 1997–1998. He joined the Faculty of the University of Illinois at Urbana-Champaign in 1999 and is currently an Assistant Professor at the Coordinated Science Laboratory at the University. His research interest include coding and information theory and their application to communication systems.

Since 1999, Dr. Koetter has served as Associate Editor for Coding Theory and Techniques for the IEEE TRANSACTIONS ON COMMUNICATIONS. In January 2000, he started a term as Associate Editor for Coding Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY. He received an IBM Invention Achievement award in 1997 and an NSF CAREER award in 2000.

G. David Forney, Jr. (S'59–M'61–F'73) received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1961, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1963 and 1965, respectively.

From 1965 to 1999, he was with the Codex Corporation, which was acquired by Motorola, Inc. in 1977, and its successor, the Motorola Information Systems Group, Mansfield, MA. He is currently Bernard M. Gordon Adjunct Professor at MIT.

Dr. Forney was Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1970 to 1973. He was a member of the Board of Governors of the IEEE Information Theory Society during 1970–1976 and 1986–1994, and was President in 1992. He has been awarded the 1970 IEEE Information Theory Group Prize Paper Award, the 1972 IEEE Browder J. Thompson Memorial Prize Paper Award, the 1990 IEEE Donald G. Fink Prize Paper Award, the 1992 IEEE Edison Medal, the 1995 IEEE Information Theory Society Claude E. Shannon Award, the 1996 Christopher Columbus International Communications Award, and the 1997 Marconi International Fellowship. In 1998 he received an IT Golden Jubilee Award for Technological Innovation, and two IT Golden Jubilee Paper Awards. He was elected a member of the National Academy of Engineering (USA) in 1983, a Fellow of the American Association for the Advancement of Science in 1993, an honorary member of the Popov Society (Russia) in 1994, and a Fellow of the American Academy of Arts of Sciences in 1998.

Frank R. Kschischang (S'83–M'91–SM'00) received the B.A.Sc. degree with honors from the University of British Columbia, Vancouver, BC, Canada, in 1985 and the M.A.Sc. and Ph.D. degrees from the University of Toronto, Toronto, ON, Canada, in 1988 and 1991, respectively, all in electrical engineering.

He is a Professor of Electrical and Computer Engineering and Canada Research Chair in Communication Algorithms at the University of Toronto, where he has been a faculty member since 1991. During 1997–1998, he spent a sabbatical year as a Visiting Scientist at the Massachusetts Institute of Technology (MIT), Cambridge, MA. His research interests are focused on the area of coding techniques, primarily on soft-decision decoding algorithms, trellis structure of codes, codes defined on graphs, and iterative decoders. He has taught graduate courses in coding theory, information theory, and data transmission.

Dr. Kschischang is a recipient of the Ontario Premier's Research Excellence Award. From October 1997 to October 2000, he served as IEEE TRANSACTIONS ON INFORMATION THEORY Associate Editor for Coding Theory. He was a member of the technical program committee for the 1995 International Symposium on Information Theory (ISIT) held in Whistler, BC, he was Co-Chair and organizer of the 1997 Canadian Workshop on Information Theory held in Toronto, and he served as Publicity Chair for the 1998 ISIT held at MIT.

Robert J. McEliece (M'70–SM'81–F'84) was born in Washington, DC, in 1942. He received the B.S. and Ph.D. degrees in mathematics from the California Institute of Technology (Caltech), Pasadena, in 1964 and 1967, respectively, and attended Trinity College, Cambridge University, Cambridge, U.K., during 1964–1965.

From 1963 to 1978, he was employed by the California Institute of Technology's Jet Propulsion Laboratory, where he was Supervisor of the Information Processing Group, Communications Reseach Section, from 1971 to 1978. From 1978 to 1982, he was Professor of Mathematics and Research Professor at the Coordinated Science Laboratory, University of Illinois, Urbana-Champaign. Since 1982, he has been on the Faculty at Caltech, where he is now the Allen E. Puckett Professor of Electrical Engineering. From 1990 to 1999, he served as Executive Officer for Electrical Engineering at Caltech. He has been a Consultant in the Communications Research Section of Caltech's Jet Propulsion Laboratory since 1978. His research interests include deep-space communication, communication networks, coding theory, and discrete mathematics.

Daniel A. Spielman was born in Philadelphia, PA, in 1970. He received the B.A. degrees in mathematics and computer science from Yale University, New Haven, CT, in 1992 and the Ph.D. degree in mathematics from the Massachusetts Institute of Technology, Cambridge, in 1995.

He spent 1995–1996 as a Postdoc in computer science at the University of California at Berkeley, where he was supported by an NSF Mathematical Sciences Postdoctoral Fellowship. Since 1996, he has been an Assistant Professor of Mathematics at the Massachusetts Institute of Technology.

Dr. Spielman was awarded a CAREER award from the NSF in 1997 and a Research Fellowship from the Alfred P. Sloan Foundation in 1998. His Ph.D. dissertation won the Association for Computing Machinery's Doctoral Dissertation Award.