Extracting Hidden Information from Knowledge Networks

Sergei Maslov[1,2] and Yi-Cheng Zhang[2]

[1]*Department of Physics, Brookhaven National Laboratory, Upton, New York 11973*

[2]*Institut de Physique Théorique, Université de Fribourg, CH-1700, Fribourg, Switzerland*

(Received 6 April 2001; published 27 November 2001)

We develop a method allowing us to reconstruct individual tastes of customers from a sparsely connected network of their opinions on products, services, or each other. Two distinct phase transitions occur as the density of edges in this network is increased: Above the first, macroscopic prediction of tastes becomes possible; while above the second, all unknown opinions can be uniquely reconstructed. We illustrate our ideas using a simple Gaussian model, which we study using both field-theoretical methods and numerical simulations. We point out a potential relevance of our approach to the field of bioinformatics.

DOI: 10.1103/PhysRevLett.87.248701 PACS numbers: 89.75.Hc, 89.20.Hh, 89.65.Gh, 89.70.+c

Mainstream economics maintains that human tastes reflected in consumer preferences are sovereign, i.e., not subject to discussion or study. It postulates that consumer's choice of products or services is the outcome of a complete and thorough optimization among all possible options, and, therefore, his/her satisfaction cannot be further improved. Such a doctrine, though often challenged from both within [1] and outside economics, is still dominant. However, recently many business practitioners started to exploit the affinity in people's tastes in order to predict their personal preferences and come up with individually tailored recommendations. Our basic premise is that people's consumption patterns are not based on the complete optimization over all possible choices. Instead, they constitute just a small revealed part of the vast pool of "hidden wants." These hidden wants, if properly exploited, can lead to much better matches between people and products, services, or other people. In the economy of the past such opportunities were hardly exploitable. Things have changed in the course of the current information revolution, which both connected people on an unprecedented scale and allowed for easy collection of the vast amount of information on customer's preferences. In just a few years the Internet has already changed much of our traditional perceptions about human interactions, both commercial and social. We believe that technical advances in wireless and other network interfaces are imminent of being able to capture the necessary information virtually free and to put this theory to use.

Our aim is to predict yet unknown individual consumer preferences, based on the pattern of their correlations with already known ones. Predictive power obviously depends on the ratio between the known and yet unknown parts. When the fraction of known opinions p is too small, only occasional predictions are possible. When it surpasses the first threshold, which we refer to as p_1, almost all unobserved preferences acquire some degree of predictability. Finally, for p above the second higher threshold p_2, all these unobserved preferences can be uniquely reconstructed. In what follows we describe a simple model of how customer's opinions are formed and spell out in some details basic algorithms allowing for their prediction.

To make this discussion somewhat less abstract let us consider a *matchmaker* or an advisor service which already exists on many bookselling websites that personally recommends new books to each of their customers. In order for such recommendation to be successful one needs to assume the existence of some "hidden metrics" in the space of reader's tastes and book's features. In other words, the matchmaking is possible only if opinions of two people with similar tastes on two books with similar features are usually not too far from each other. In this work we use the simplest realization of this hidden metrics. We assume that each reader is characterized by an M-dimensional array $\mathbf{r} = (r^{(1)}, r^{(2)}, \ldots, r^{(M)})$ of his/her tastes in books, while each book has the corresponding list of M basic "features" $\mathbf{b} = (b^{(1)}, b^{(2)}, \ldots, b^{(M)})$ [1]. An opinion of a reader on a book is given simply by an overlap (scalar product) Ω of reader's vector of tastes and book's vector of features: $\Omega = \mathbf{r} \cdot \mathbf{b} = \sum_{\alpha=1}^{M} r^{(\alpha)} b^{(\alpha)}$. The matchmaker has some incomplete knowledge about opinions of his customers on the books they have read, and he uses it to reconstruct yet unknown opinions (overlaps) and to recommend books to its customers.

The *central* position of our matchmaker with respect to its customers makes its services dramatically different from those of the so-called "smart agents" [2], whose goal is to anticipate and predict tastes of their individual owners. Indeed, the scope of recommendations of a smart agent is severely limited by the fact that each of them serves its own master, so that others would not cooperate. On the other hand, our matchmaker is a completely neutral player in an economic game, who is able to synergistically use the knowledge collected by all players/agents to everybody's advantage (including his own).

The information about who-read-what is best visualized as a bipartite undirected graph in which vertices corresponding to readers are connected by edges to vertices corresponding to books each of them has read and reported opinion to the matchmaker. Similar graphs (or networks)

were recently drawn to the center of attention of the statistical physics community [3–5] under a name "small world networks." For example, statistical properties of a bipartite graph of movie actors connected to films they appeared in were studied in [3,4], while that of scientists and papers which they coauthored were studied in [5]. In this Letter we go beyond empirical studies or simple growth models of such graphs. The new feature making the graphs introduced in this work richer than ordinary undirected graphs is that in our graphs each vertex has a set of M "hidden" internal degrees of freedom. Consequently, each edge carries a real number Ω, representing the similarity or overlap between these internal degrees of freedom on two vertices it connects. In our case this number quantifies the matchmaker's knowledge of an opinion that a given customer has on a given product. Therefore, we would refer to such graphs as *knowledge or opinion networks*.

In the most general case any two vertices in the knowledge network can be connected by an edge. It is realized, for instance, if vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ stand for strings of individual "interests" in a group of N people. The overlap $\Omega_{ij} = \mathbf{r}_i \cdot \mathbf{r}_j$ measures the similarity of interests for a given pair of people and can be thought of as the "quality of the match" between them. The matchmaker's goal is to analyze this information and to recommend to a customer i another customer j, whom he has not met yet, and who is likely to have a large positive overlap with his/her set of interests. Mutual opinions can be conveniently stored in an $N \times N$ *symmetric* matrix of scalar products $\hat{\Omega}$. In the above case any element of this matrix can be in principle "reported" to the matchmaker. Different restrictions imposed on this most general scenario describe other versions of our basic model such as the following: (1) An advisor service recommending N_b products to N_r customers (e.g., our model of books and readers from the introduction). In this case the square matrix $\hat{\Omega}$ has $N_r + N_b$ rows and columns, while all entries known to the matchmaker are restricted to the $N_r \times N_b$ rectangle, corresponding to opinions of customers on products. (2) A real matchmaking service recommending N_m men and N_w women to each other. Here we assume that each man and woman can be characterized by two M-dimensional vectors: the first one is the vector $\mathbf{q}$ of his/her own "qualities," while the second one $\mathbf{d}$ represents the set of his/her "desires," i.e., desired ideal qualities that he/she is seeking in his/her partner. The opinion of a person i on a person j is then given by a scalar product $\mathbf{d}_i \cdot \mathbf{q}_j$, while the opposite opinion has in general a completely different value $\mathbf{d}_j \cdot \mathbf{q}_i$. The full $(2N_m + 2N_w) \times (2N_m + 2N_w)$ overlap matrix is still symmetric but only two small sectors, containing $N_m \times N_w$ elements each, are accessible to the matchmaker.

With a small modification this last scenario can be applied to a completely different problem, namely that of physical interactions in a set of biological molecules such as proteins. It is known that high specificity of such interactions is achieved by the virtue of the "key-and-lock" matching of features on their surfaces. Given the space of possible shapes of locks and keys, each molecule can be described by two vectors $\mathbf{l}_i$, $\mathbf{k}_i$ of 0's and 1's which determine which keys and locks are present on its surface. Provided that the key k^α uniquely fits the lock l^α, the strength of the interaction between these two molecules is determined by $\Omega_{ij} = \mathbf{k}_i \cdot \mathbf{l}_j + \mathbf{k}_j \cdot \mathbf{l}_i$.

In the rest of the paper we concentrate only on the most general nonbipartite case of an $N \times N$ matrix of overlaps of interests in a group of N customers and leave other more restricted situations for future work [6]. The matchmaker always has only partial and noisy information about the matrix $\hat{\Omega}$ due to several factors: (1) First and most importantly, the matchmaker knows only some of the opinions Ω_{ij} of his customers on each other, which he uses to guess the rest. (2) In real life the overlap could never be precisely measured. In the simplest case of an extremely narrow information channel customers report to the matchmaker only the *sign* of their overlap with other customers. One can also imagine a somewhat wider channel, where the matchmaker asks his customers to rate their satisfaction by a grade system, the finer the better. (3) The loss of information due to a narrow channel between the matchmaker and its customers can be further complicated by a random noise in reporting, which would inevitably be present in real life situations. Indeed, we are far from assuming that the scalar product of tastes and features completely determines the customer satisfaction with a product, or that similarity of interests is all that matters when two people form an opinion about each other. One should always leave room for an idiosyncratic reaction, which does not result from any logical weighting of features. Our hope is that strong mutually reinforcing correlations due to the redundance of information stored in an idealized matrix $\hat{\Omega}$ would manifest themselves in a large enough group of customers even when they are masked by a substantial amount of idiosyncratic noise. In principle, all these three sources of noise and partial information are present simultaneously. However, in this work we will treat them separately and restrict ourselves only to the case where the matchmaker knows the *exact* values of all overlaps reported to him. It is easy to see how correlations between matrix elements allow the matchmaker to succeed in his goal of prediction of yet unknown overlaps. For example, the known values of $\Omega_{12} = \mathbf{r}_1 \cdot \mathbf{r}_2$ and $\Omega_{23} = \mathbf{r}_2 \cdot \mathbf{r}_3$ somewhat restrict the possible mutual orientation of vectors $\mathbf{r}_1$ and $\mathbf{r}_3$, and, therefore, contain information about the value of the yet unknown overlap Ω_{23}. Below we will demonstrate that the predictability of an overlap between two points that are already connected by a chain of known overlaps of length L is proportional to $M^{-(L-1)/2}$ and, therefore, exponentially decays with L for $M > 1$. Hence, an appreciable prediction becomes possible only when two points are connected by exponentially many mutually reinforcing paths.

The amount of information collected by the matchmaker on its customers can be conveniently characterized by either the number K or the density $p = 2K/N(N-1)$ of known overlaps among all $N(N-1)/2$ off-diagonal elements of the matrix. For very small K all edges of the knowledge network are disconnected and no prediction is possible. As more and more edges are *randomly* added to the network, the chance that a new edge would join two previously connected points, i.e., the probability to form a loop in the network, increases. It is exactly in this situation the matchmaker had some predictive power about the value of the new overlap before it was observed. However, this excess information would disappear in the thermodynamic limit $N \to \infty$ until the density of edges reaches the first threshold $p_1 = 1/(N-1)$. This threshold is nothing else but a percolation transition, above which the giant connected component (GCC) appears in a random graph. For $p > p_1$ the fraction of nodes in the GCC rapidly grows, exponentially approaching 100%. It means that already for a moderate ratio p/p_1 almost every new edge added to the graph would join two previously connected points. This transition would also manifest itself in the behavior of the entropy of the joint probability distribution of unknown overlaps [6].

One has to remember though that the predictive power of the matchmaker is exponentially small for long loops. That means that while the typical diameter of the graph is still large, the loop correlation is too weak to significantly bias most of the unknown overlaps. The reliable prediction becomes possible only for much higher values of p. Let us calculate p_2—the point of the second phase transition, above which the values of *all* unknown overlaps are completely determined by the information contained in known ones. Using a geometrical language at this point the knowledge network undergoes a "rigidity percolation" phase transition, at which relative orientations of vectors $\mathbf{r}_i$ become fixed. Such transition is possible only for $N > M$ since only in this case $\hat{\Omega}$ contains *redundant* information about components of all vectors $\mathbf{r}_i$. The position of the second phase transition p_2 can be determined by carefully counting the degrees of freedom. For $N > M$ the overlap matrix $\hat{\Omega}$ has very special spectral properties: it has precisely $N - M$ zero eigenvalues, while the remaining eigenvalues are strictly positive. An easy way to demonstrate this is to recall that the overlap matrix can be written as $\hat{\Omega} = \hat{R}\hat{R}^\dagger$, where $\hat{R}$ is the $N \times M$ rectangular matrix formed by vectors $r_i^{(\alpha)} = R_{i\alpha}$. The singular value decomposition (SVD) technique allows one to "diagonalize" $\hat{R}(N > M)$, that is, to find an $M \times M$ orthogonal matrix $\hat{V}$ ($\hat{V}\hat{V}^\dagger = 1$), an $M \times M$ positive diagonal matrix $\hat{D}$, and an $N \times M$ matrix $\hat{U}$ formed by M orthonormal N-dimensional vectors, such that $\hat{R} = \hat{U}\hat{D}\hat{V}$. Now it is easy to see that $\hat{\Omega} = \hat{U}\hat{D}^2\hat{U}^\dagger$ has precisely M *positive* eigenvalues equal to squares of the elements of the diagonal matrix $\hat{D}$, and $N - M$ zero eigenvalues. The number of degrees of freedom of $\hat{\Omega}$ is equal to the NM degrees of freedom of R minus $M(M-1)/2$ of the "gauge" degrees of freedom of the orthogonal matrix V, which have no influence on elements of $\hat{\Omega}$. Once the number of known elements K exceeds the total number of degrees of freedom of $\hat{\Omega}$, the remaining unknown elements of $\hat{\Omega}$ can be in principle reconstructed. Therefore, the second phase transition happens at

$$p_2 = \frac{M(2N - M + 1)}{N(N-1)} \simeq 2M/N. \qquad (1)$$

Here the $\simeq$ sign corresponds to the limit $N \gg M$.

Practically, however, in order to calculate the set of unknown overlaps one needs to solve a system of nonlinear equations with a huge number of unknown variables, which is a daunting task. To this end we came up with a simple and efficient iterative numerical algorithm that uses the special spectral properties of $\hat{\Omega}$: (1) Construct the initial approximation $\hat{\Omega}_a$ to $\hat{\Omega}$ by substituting 0 for all its *unknown* elements. (2) Diagonalize $\hat{\Omega}_a$, and construct the matrix $\hat{\Omega}'_a$ by keeping the M largest (positive) eigenvalues and eigenvectors of $\hat{\Omega}_a$, while setting the remaining $N - M$ eigenvalues to zero. (3) Construct the new refined approximate matrix $\hat{\Omega}_a$ by copying all unknown elements from $\hat{\Omega}'_a$, while resetting the rest to their *exactly known values*. (4) Go to the step (2). As shown in Fig. 1 for $p > p_2$, $\hat{\Omega}_a$ converges to $\hat{\Omega}$ exponentially fast in the number of iterations n. Numerical simulations also indicate that the rate of this exponential convergence scales as $(p - p_2)^2$ above the second phase transition (see the inset of Fig. 1).

Below p_2 this algorithm performs rather poorly and the error may even grow with the number of iteration steps. This is to be expected since in this region there is more

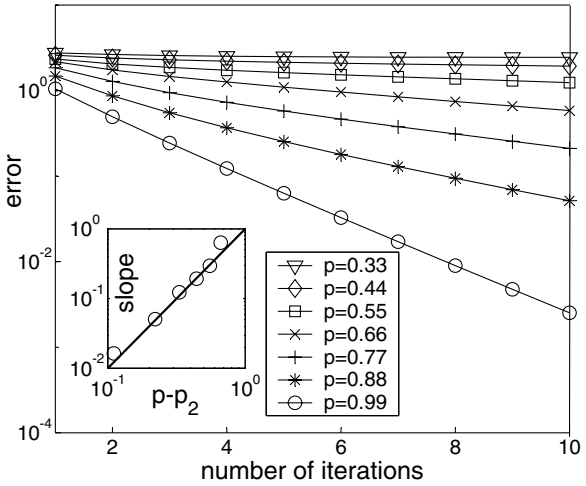

FIG. 1. The average error in the value of unknown matrix elements of $\hat{\Omega}$ as a function of the number of iterations. All $r_i^{(\alpha)}$ are independent Gaussian random numbers. The parameters of the model are $M = 9$, $N = 50$, corresponding to $p_2 = 0.34$. The inset shows the scaling of an exponential convergence rate as a function of $p - 0.34$. The solid line has the slope 2.

than one solution for the $\hat{\Omega}$, consistent with a set of constraints, imposed by K known matrix elements. While our iterative algorithm always converges to *one* of such solutions, barring an unlikely accident, this solution is far from the set of "true" values of unknown matrix elements. In this situation the best thing that a matchmaker can do is to calculate the average value $\langle \Omega_{pq} \rangle$ of each unknown element in the ensemble of all matrices, consistent with a given set of K constraints. We have succeeded in estimating $\langle \Omega_{pq} \rangle$ analytically. This calculation involves rather heavy algebra and will be reported elsewhere [6].

In the above discussion the parameter M was treated as fixed and known property of the system. However, in real life one usually does not know *a priori* the number of relevant components of an idealized vector of tastes or features. Here we want to propose a criterion on how to optimally choose it. If the number of known overlaps K is small, it would be useless to try to model the matrix using a high-dimensional space of tastes. Indeed, all the free play allowed by a large M would not give the matchmaker much of a prediction power anyway. This leads us to a conjecture that the optimal way for a matchmaker to select an effective number of internal degrees of freedom $M_{\rm eff}$ is to do it in such a way that the system is balanced precisely at or near the critical threshold p_2. In other words, given K and N one should solve the equation $NM_{\rm eff} - M_{\rm eff}(M_{\rm eff} - 1)/2 = K$ to find $M_{\rm eff} = [N + 1/2 - \sqrt{(N + 1/2)^2 - 2K}] \cong K/N$.

Finally, we introduce a particularly simple analytically tractable example of a knowledge network, where each component r_i^α of a hidden vector $\mathbf{r}_i$ is *independently* drawn from a normal distribution. The joint probability distribution $P(\hat{\Omega})$ of all $N(N + 1)/2$ elements of the (symmetric) overlap matrix $\hat{\Omega}$ is then given by a multidimensional integral $P(\hat{\Omega}) = \int\int \cdots \int \prod_{i,\alpha} (dr_i^{(\alpha)}/\sqrt{2\pi}) \times \exp[-\sum_{i,\alpha} (r_i^{(\alpha)})^2/2] \prod_{i \le j} \delta(\Omega_{ij} - \sum_{\alpha=1}^{M} r_i^{(\alpha)} r_j^{(\alpha)})$. Using the standard integral representation for the δ function, $\delta(x) = \int_{-\infty}^{\infty} \exp(i\lambda x)\, d\lambda/(2\pi)$, and calculating exactly the path integral, now quadratic in $r_i^{(\alpha)}$, one arrives at a remarkably elegant and compact expression [7]:

$$P(\hat{\Omega}) = \int\int \cdots \int_{-\infty}^{\infty} \prod_{i \le j} \left(\frac{d\lambda_{ij}}{2\pi} \exp(i\lambda_{ij}\Omega_{ij}) \right) \times \det(\hat{1} + i\hat{\Lambda})^{-M/2}. \quad (2)$$

The matrix $\hat{1}$ is the $N \times N$ unity matrix, while $\hat{\Lambda}$ is a symmetric matrix with elements $2\lambda_{ii}$ on the diagonal and λ_{ij} off the diagonal. This expression is the multidimensional Fourier transform of the joint probability distribution $P(\hat{\Omega})$, so that $\Phi(\hat{\Lambda}) = \det(\hat{1} + i\hat{\Lambda})^{-M/2}$ is nothing else but the *generating function* of this distribution. As usual, Taylor expansion of the generating function in powers of λ_{ij} around $\hat{\Lambda} = 0$ allows one to calculate any imaginable correlation between integer powers of Ω_{ij}. It is more convenient to work with irreducible correlations, generated by the Taylor expansion of $\phi(\hat{\Lambda}) = \ln(\Phi(\hat{\Lambda})) = -(M/2)\ln[\det(\hat{1} + i\hat{\Lambda})] = -(M/2)\mathrm{Tr}[\ln(\hat{1} + i\hat{\Lambda})]$. A surprising consequence of the above exact expression for $\phi(\hat{\Lambda})$ is that *all* irreducible correlations of matrix elements are proportional to M. In particular, the expansion $\phi(\hat{\Lambda}) = (M/2)\sum_{L=1}^{\infty} \mathrm{Tr}[(-i\hat{\Lambda})^L]/L$ allows one to calculate any correlation of the type $\langle\langle \Omega_{i_1 i_2} \Omega_{i_2 i_3} \cdots \Omega_{i_{L-1} i_L} \Omega_{i_L i_1} \rangle\rangle = M$, corresponding to a given non-self-intersecting loop on the network. The presence of such cyclic correlations indicates that signs of matrix elements are weakly correlated. Taking into account that each $|\Omega_{ij}| \sim \sqrt{M}$ and using scaling arguments it is straightforward to demonstrate that the predictability of one of the overlaps in the loop of length L based on the knowledge of others scales as $M^{-(L-1)/2}$.

In this Letter we have described a general framework allowing one to predict elements from the unobserved part of a knowledge network based on the observed part. Prediction power was shown to strongly depend on the ratio between these two parts. While our original motivation was to model a commercial matchmaking service in the internet age, the implications go well beyond. We point out that our general framework, developed for knowledge networks, could also be of much importance in the field of bioinformatics, where cross correlations, mutual interactions, and functions of large sets of biological entities such as proteins, DNA binding sites, etc., are only partially known. It is conceivable that a similar approach applied to, e.g., a large matrix of protein-protein interactions [8] would prove to be fruitful.

We have benefited from conversations with T. Hwa, M. Marsili, C. Tang, Y. Yu, and A. Zee. Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy. This work was sponsored in part by Swiss National Foundation under Grant No. 20-61470.00.

[1] See, e.g., K. Lancaster, J. Political Economy **74**, 132–155 (1966).
[2] For a discussion on pros and cons of smart agents see, e.g., www.shirky.com/writings/bots.html
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] A.-L. Barabasi and R. Albert, Science **286**, 509 (1999).
[5] M. E. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001); cond-mat/0011144.
[6] S. Maslov and Y.-C. Zhang (to be published).
[7] As was demonstrated by J. Wishart in 1928 [see, e.g., S. Wilks, *Mathematical Statistics* (John Wiley, New York, 1967)] for $M > N$ the following integral can be calculated exactly to give $P(\hat{\Omega}) = \det(\hat{\Omega})^{(M-N-1)/2} \times \exp[-\mathrm{Tr}(\hat{\Omega})/2]/\mathrm{norm}$. However, for $M < N$, which is our case, this problem turns out to be much harder.
[8] P. Uetz *et al.,* Nature (London) **403**, 623 (2000); T. Ito *et al.,* Proc. Natl. Acad. Sci. U.S.A. **98**, 4569 (2001).