Centre de Referència en Economia Analítica

Barcelona Economics Working Paper Series

Working Paper nº 64

Optimal Information Transmission in Organitzations: Search and Congestion

Àlex Arenas, Antonio Cabrales, Leon Danon, Albert Díaz-Guilera, Roger Guimerà and Fernando Vega-Redondo

July, 2003

Optimal Information Transmission in Organizations: Search and Congestion*

Àlex Arenas[†], Antonio Cabrales[‡], Leon Danon [§],
Albert Díaz-Guilera[¶], Roger Guimerà[‖], Fernando Vega-Redondo[**]

July 2003
Barcelona Economics WP n^{o}64

Abstract

We propose a stylized model of a problem-solving organization whose internal communication structure is given by a fixed network. Problems arrive randomly anywhere in this network and must find their way to their respective "specialized solvers" by relying on local information alone. The organization handles multiple problems simultaneously. For this reason, the process may be subject to congestion. We provide a characterization of the threshold of collapse of the network and of the stock of floating problems (or average delay) that prevails below that threshold. We build upon this characterization to address a design problem: the determination of what kind of network architecture optimizes performance for any given problem arrival rate. We conclude that, for low arrival rates, the optimal network is very polarized (i.e. star-like or "centralized"), whereas it is largely homogenous (or "decentralized") for high arrival rates. We also show that, if an auxiliary assumption holds, the transition between these two opposite structures is sharp and they are the only ones to ever qualify as optimal.

*The authors would like to thank L. A. N. Amaral, X.Guardiola, R. Monasson, C.J. Pérez, M. Sales and seminar audiences for helpful comments and discussions. This work has been supported by the DGES of the Spanish Government, Grants No. PPQ2001-1519, No. BFM2000-0626, No. BEC2000-1029 and No. BEC2001-0980, as well as by the EC-Fet Open project IST-2001-33555. Roger Guimerà also acknowledges financial support from the Generalitat de Catalunya.

[†]Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili. e-mail: aarenas@etse.urv.es.

[‡]Departament d'Economia i Empresa, Universitat Pompeu Fabra. e-mail: antonio.cabrales@econ.upf.es.

[§]Departament de Física Fonamental, Universitat de Barcelona. e-mail: ldanon@ffn.ub.es.

[¶]Departament de Física Fonamental, Universitat de Barcelona. e-mail: albert@ffn.ub.es.

[‖]Department of Chemical Engineering, Northwestern University. e-mail: rguimera@northwestern.edu.

[**]Departament de Fonaments de l'Anàlisi Econòmica, Universitat d'Alacant and Departament d'Economia i Empresa, Universitat Pompeu Fabra. e-mail: vega@merlin.fae.ua.es.

1 Introduction

Efficient information transmission is one of the most pressing problems faced by organizations, say firms. This is specially important in modern economies, for at least two reasons. One is that more firms now are pure knowledge-based outfits (think of large engineering, consulting, research and development or financial services enterprises). The other is that with an ever increasing stock of knowledge, most individuals cannot be reasonably expected to master significant fractions of that knowledge.

Thus, the amount of available knowledge, plus the limitations inherent to the human mind, make knowledge specialization a necessity. Yet there is another limitation that comes with specialization. We not only ignore certain things, but also ignore who knows them. Without this limitation, it would be simple to deal with information transmission within organizations (barring incentive problems, from which we abstract). Suppose anybody in an organization had a problem she could not solve. She would only need to contact the expert in the topic, who would then deal with it. Some classes of problems are, arguably, simple enough that this mode of information transmission would be sufficient. This paper deals with classes of problems where being aware of the knowledge sets of others is a scarce resource.

In this context, we explore what is the most efficient form of organizing communication. The organization is modelled as a network, whose objective is to solve problems. The individuals are the nodes of this network and they have the ability to solve a particular class of problems. New problems originate at randomly chosen nodes, and for every problem there is another, independently chosen, node within the organization who can solve it. The (mutual) knowledge of two individuals about each other's abilities are the links of this network. That is, individuals only know whether they can solve a problem that arrives to them (either because the problem originates with them, or because another member of the organization handed it to them), or whether any of their directly linked neighbors can do it. The search algorithm that routes information through the organization can only use that knowledge. Our aim is to find the best way to connect the nodes, given a fixed number of links and an algorithm with purely local knowledge.

The fundamental relationship we uncover is a trade-off between decreasing the average distance between nodes and the countervailing effect on performance induced by problem overload and congestion. If congestion were not an issue, the optimal organizational structure would be very polarized. If one node were connected with all the rest, and that node were the only one with which the others were connected (a star-like organization), any problem could reach its solution in, at most, two steps. The number of links required for this would be one less than the number of nodes. The drawback of this organizational form is that it would collapse when the average number of problems arriving to an organization per period were larger than the number of problems the center could handle per period.

Motivated by these considerations, our first contribution is to solve (given any organizational structure) for the smallest rate of problem generation such that the average stock of unsolved pending problems in the organization diverges to infinity, that is, the network collapses. Furthermore, for arrival rates of new problems that are smaller than this critical value, we determine its average stock of floating problems. This stock, in turn, is directly related to the average length of time that each problem spends in the organization. It can, thus, be interpreted as a measure of the "quality of the (problem-solving) service" that the organization provides. Using this characterization of the average delay, we then turn to considering what is the optimal organizational form that minimizes the delay. For low rates of problem arrival, we conclude that it is a polarized (star-like or "centralized") network, whereas for high ones it is an homogenous (or "decentralized") structure. Making an auxiliary assumption, we also find that the degree of polarization of the optimal network varies monotonically (in a weakly non-decreasing fashion) with the rate of problem arrival. In fact, our substantially stronger finding in this respect is that that transition between the extreme kinds of network (i.e. the polarized and the homogenous) is abrupt, with only these two structures ever arising as optimal.

As indicated, the latter results depend on an auxiliary assumption that pertains to how the set of admissible networks translates into the corresponding set of so-called "betweenness" (roughly, the betweenness of a node is a measure of the centrality bestowed on it by the search protocol). Specifically, it is posited that the lower frontier of the "betweenness possibility set" displays the same curvature throughout (i.e. concave or convex). Even though this assumption is plausible, it is very hard to check directly. This is why we conclude the paper by exploring an array of different specific environments where the design issue is addressed numerically. In all of them, the optimal network architectures are found to display the behavior predicted by the theoretical analysis.

The paper is organized as follows. Section 2 describes the model. Section 3 carries out the analysis by completing, in turn, the following steps: the study of a benchmark setup without congestion (Subsection 3.1), the analytical characterization of the collapse threshold (Subsection 3.2), an analogous task for the problem load (Subsection 3.3), and the organizational design problem (Subsection 3.4). Section 4 discusses the related literature. Section 5 summarizes and discusses some avenues for further research.

2 The model

Our organization will be modelled as a network, or more precisely by an undirected graph. In this graph, the nodes are the individual members of the organization. Let $N = \{1, 2, ..., n\}$ be the set of all individual nodes. Each individual can solve some specific class of problems. A link between two nodes i and j implies that both individuals know the set of problems that the other individual in the pair can solve. Formally, for each pair of nodes i and j, we define

$g_{ij} \in \{0,1\}$. The condition $g_{ij} = 1$ is taken to imply that the two nodes are linked, whereas $g_{ij} = 0$ implies that the two nodes are not linked. Since the graph is undirected, $g_{ij} = 1$ if and only if $g_{ji} = 1$. Let $\Gamma = \{N, (g_{ij})_{i,j=1}^{n}\}$ be a given network. Then, the set of neighbors of any given agent $i \in N$, denoted by N_i, is given by $N_i = \{j \in N : g_{ij} = 1\}$.

The mission of this organization is to solve problems. At each point in time, modelled continuously, problems make their first appearance in an organization at an independent rate ρ at each node. Each problem starting at $i \in N$ has an "address" indicating the node k where it is to be solved. We, thus, implicitly assume that individual knowledge is sufficiently specific that each problem can be solved by only one person.[1] Let us then refer to "problem k" as any problem that can be solved only at node k. Typically, of course, k will be different from the node where it arrives.

We now have to define the rules by which the problem travels through the organization. If the node where the problem arrives, either at the beginning of the process or at some intermediate step, can solve it, then it will do so and the problem disappears from the organization. We will now specify the rules determining further travel, when the node which receives the problem cannot solve it. But first notice that there may be several problems "waiting" at node i, at any point in time. Not all of them may be chosen to travel further at one particular time. The rules through which "queues" are managed will be specified in section 3.1. We will now explain how problems that are chosen to travel further "decide" a destination. Denote by p_{ij}^{k} the probability with which a problem k being at node i will go to node j if chosen to be sent forward.[2]

Once a problem k is at (faced by) node i, one the following two alternative rules are applied:

- If $k \in N_i$, the problem is sent to k with $p_{ik}^{k} = 1$ and it is solved immediately.
- If $k \notin N_i$, the problem is sent to some $j \in N_i$ with some probability p_{ij}^{k}. (Of course, $\sum_{j \in N_i} p_{ij}^{k} = 1$.)

Any problem proceeds as above until solved. The first rule should not be controversial. The second rule assumes that the knowledge that individuals can use to route problems is the identity of their neighbors, and the final destination. This implicitly allows them to have the underlying network geography in mind, but not exploit the knowledge of what is the current state of congestion (even at the level of first neighbors). Such an assumption is taken here for convenience, and we presume that little of interest would be changed by relaxing it.

[1] In the literature review we discuss alternative approaches.

[2] Since the problem is supposed not yet to be solved, we are implicitly assuming that $i \neq k$. However, if we had $i = k$, it is formally convenient to simply make the corresponding travel probabilities uniformly zero, i.e. $p_{kj}^{k} = 0$ for all $k \in N$.

The network combined with the protocol that guides the problems lead to a collection of communication (pseudo-stochastic) matrices

$$\{P^k \equiv (p^k_{ij})_{i,j\in N}\}_{k\in N}. \tag{1}$$

These matrices define the stochastic process that governs the steps (or direction) followed by the each problem k. In line with the previous discussion, they are assumed to display the following features:

$$\begin{aligned} p^k_{ij} &= 0 \quad \text{if} \quad j \notin N_i \\ p^k_{ik} &= 1 \quad \text{if} \quad k \in N_i \\ p^k_{kj} &= 0 \quad \forall j \in N. \end{aligned}$$

We may compute, for each $r \in \mathbb{N}$:

$$q^k_{ij}(r) = \sum_{l_1,l_2,\ldots,l_{r-1}} p^k_{il_1} p^k_{l_1 l_2} \cdots p^k_{l_{r-1}j}$$

as the probability of a problem k currently in i to be in node j after r steps. Or, using matrix notation, we may simply define $Q^k(r)$ as the matrix whose ijth element is $q^k_{ij}(r)$ so that:

$$Q^k(r) = (P^k)^r = P^{k\,(r \text{ times})} \cdots P^k$$

To be sure, note that the above probabilities only govern the direction of movement of the packages, but not necessarily the time they spend unsolved. To address the latter, we need to superimpose on the above "congestion-blind" formulation the processing delays which may impede swift movement of packages across nodes in the presence of waiting queues.

3 Analysis

3.1 Steady-state analysis and the threat of collapse

Now, let us return to the case which has motivated our approach, where each agent/node has limited processing capability. Specifically, we assume that the nodes behave as *queues.* This means that they have unlimited storage capacity but process problems, in expected terms, at a constant rate per instant of time, which we normalize to unity. Thus, under the maintained assumption of stationarity, the number of pending problems standing in a queue behaves like an infinite-state Markov process and the arrivals and departures from each node i follow Poisson processes. As long as the fluctuations have finite variance, the overall process displays well-defined steady state probabilities and averages.

Thus suppose that the process reaches a steady state and let us describe its characteristics. Denote by a_{ij}^k the stationary arrival rate to node j of problems which appeared in the network at node i with destination k, and let δ_{ij}^k stand for the stationary departure rate of problems from node j of problems which appeared in the network at node i with destination k. Then, since the arrival rate to a node is the sum of the arrival rate from the outside of the system (new

problems) plus arrival rates from other nodes we have:[3]

$$a_{ij}^k = \begin{cases} \frac{\rho}{n-1} + \sum_{l=1}^{n} \delta_{il}^k p_{lj}^k, & \text{when } j \neq k \\ 0, & \text{when } j = k \end{cases} \tag{2}$$

The second line is zero, since we assume that problems that reach their destination get solved, so they do not get added to the queue. But given that in steady state all problems that arrive to a node eventually depart from it in finite time, we must have that $a_{ij}^k = \delta_{ij}^k$ for all i, j, k and therefore:

$$a_{ij}^k = \begin{cases} \frac{\rho}{n-1} + \sum_{l=1}^{n} a_{il}^k p_{lj}^k, & \text{when } j \neq k \\ 0, & \text{when } j = k. \end{cases} \tag{3}$$

Let R^k be a diagonal matrix such that $r_{ij}^k = 1$ for $i = j \neq k$ and $r_{ij}^k = 0$ otherwise. Now, making $A^k \equiv (a_{ij}^k)_{i,j \in N}$, we can write the equations (3) in matrix form as follows:

$$\begin{aligned} A^k &= \frac{\rho}{n-1} R^k + A^k P^k R^k \\ A^k &= \frac{\rho}{n-1} R^k (I - P^k R^k)^{-1} \end{aligned}$$

[3] The queuing network considered here is closely related to what is known in the Operations Research literature as a multi-class Jackson network (see e.g. Chao, Miyazawa and Pinedo 1999). These networks are known to generate an ergodic Markov process whose invariant distribution is a *product measure*. This property is also satisfied in our case and permits analyzing the flow of problems faced by each node as a composition of independent Poisson processes. Consequently, the arrival rates from different sources can be made to add up to a combined arrival rate, as postulated in (2).

In order to interpret the induced arrival rates, let us consider a (fictitious) scenario, in which time is discrete and the number of nodes visited by a problem is equivalent to the time it spends in the network. That is, all problems arriving to a node on any given period are always dispatched prior to entering the following period without delay. Further assume, in order to fix ideas, that, for *every* k and i, a problem k is created in i with probability one at each period. Then, the probability $q_{ij}^k(r)$ defined at the end of section 2 can be trivially reinterpreted as the probability that, at any given time $t(\geq r)$, there is a problem k which originated r periods ago in node node i that is currently faced by node j. With this interpretation in mind, the expression

$$b_{ij}^k \equiv \begin{cases} \sum_{r=0}^{\infty} q_{ij}^k(r), & \text{when } j \neq k \\ 0, & \text{when } j = k \end{cases}$$

can be viewed as the limiting (or steady-state) *expected* number of problems k which arose in i sometime in the past and are currently passing through j at some "distant" period t.[4] Let B^k denote the matrix $(b_{ij}^k)_{i,j\in N}$ for any given k. Then, compactly, we may write in matrix form:

$$B^k = \sum_{r=0}^{\infty} Q^k(r)R^k = \sum_{r=0}^{\infty}(P^k)^r R^k = (I - P^k)^{-1}R^k$$

[4] We have $b_{ik}^k = 0$, since we assumed that problems that reach their destination are solved immediately.

Based on these magnitudes, let us define the (algorithmic) *betweenness* of any particular node j by:

$$\beta_j \equiv \sum_{i=1}^{n} \sum_{k=1}^{n} b_{ij}^k$$

That is, we simply add over all possible origins i and destinations k. In line with the previous discussion, one can interpret β_j as the expected number of problems (of any kind, and with any origin) that are going through node j in the long run.[5] The magnitude embodied by each β_j abstracts from considerations of congestion. We will see, nevertheless, that this magnitudes bears a very strong connection with the behavior of the model, in particular concerning the arrival rates displayed in A^k.

To make this connection, we need to carry out the following derivations. Notice first that since $p_{kj}^k = 0$ for all j, we have that $R^k P^k = P^k$. Postmultiplying both matrices by R^k, this implies that:

$$-R^k P^k R^k = -P^k R^k$$

Adding R^k on both sides and then isolating the common factor R^k also on both sides we have:

$$R^k \left[I - P^k R^k\right] = \left[I - P^k\right] R^k$$

Now, premultiplying both sides by $\left[I - P^k\right]^{-1}$ and postmultiplying $\left[I - P^k R^k\right]^{-1}$

$$\left[I - P^k\right]^{-1} R^k = R^k \left[I - P^k R^k\right]^{-1}$$

[5] Note that the present notion of *betweenness* is algorithmic-based, in the sense that it is associated to the particular search protocol used by the organization. Thus, it is to be distinguished from the more usual notion of *topological betweenness* (Freeman 1977, Newman 2001), which assumes that the search algorithm at work is globally efficient and is able to identify the minimal distance paths between nodes.

so that $A^k = \frac{\rho}{n-1} B^k$. This implies that if we denote by $\alpha_j = \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ij}^k$ the total arrival rate of problems to a node (from every origin i and destination k), then

$$\alpha_j = \frac{\rho}{n-1} \beta_j \tag{4}$$

i.e. the total problem arrival rate faced by any node is proportional to its betweenness.

Recall that we have normalized the departure rate of problems from each non-idle node to 1. Under these conditions, the length of the queue is expected to grow without bound if, and only if, the expected number of problems arriving every period to the queue is larger than the expected number of problems that can be processed in each period. Therefore, relying on (4), we can formulate matters in terms of the corresponding betweenness and state that a particular node j collapses, provided no other does, iff

$$\frac{\rho}{n-1} \beta_j > 1,$$

which implies that the maximum ρ consistent with *no* node collapsing in the network is:

$$\rho_c = \frac{n-1}{\beta^*} \tag{5}$$

where $\beta^* \equiv \max_j \beta_j$ is the *maximum betweenness*.

At this point, it may be useful to provide a concrete example that naturally fits in our theoretical framework. Consider a scenario where:

(a) the probabilities p_{ij}^k that define the communication protocol of the organization are unbiased in the following sense: For all $i, j, k \in N$, such that $i \neq k$

and $k \notin N_i$,[6]

$$p_{ij}^k = \frac{1}{|N_i|}.$$

(b) For every problem k awaiting at node i, this problem is processed with independent probability equal to $\frac{1}{q_i}$, where q_i stands for the number of problems in the queue.[7]

Any scenario satisfying (a)-(b) is consistent with our maintained assumptions, i.e. its communication protocol can be described by a corresponding set of matrices as in (1) and the nodes behave as queues (they process an expected number of problems equal to unity). Notice that assumption (a) precludes the possibility that a problem is routed taking into account its final destination. This is consistent with our philosophy that the links represent the mutual knowledge of two individuals about each other's abilities. Thus, the absence of a link with k implies that individual i (with $k \notin N_i$) has no knowledge of the "best" direction of movement. In the concluding remarks we discuss what can happen when this assumption is relaxed.

3.2 Organizational performance

Assume that for all $i \in N$, $\frac{\rho}{n-1}\beta_i < 1$, that is, the expected number of arrivals to all nodes is smaller than the expected number of exit opportunities. This, as explained, averts the possibility of collapse. However, the fact that, in expected terms, the number of unsolved problems cannot grow unboundedly does not rule out the possibility that queues of positive length might persist throughout the network. To understand this intuitively, note that the (unavoidable) fluctuations that are forever present along the process induce inherently asymmetric effects on the length of queues. On the one hand, when no problems stand in the queue of a certain node, the queue can obviously become no shorter. Instead, no matter how long a queue might be, there is always positive probability that it increases even further. In heuristic terms, one could describe the basis of this asymmetry

[6] Recall that, if $k \in N_i$, it was required that $p_{ik}^k = 1$.

[7] We could easily handle non-random disciplines for problem delivery, like FIFO (First-In-First-Out). The advantage of a random discipline is that is minimizes the amount of memory needed for numerical computation (as the algorithm does not need to keep track of an order of arrival to the queue at each node). Thus, it speeds up the simulations we perform in the next section.

as follows: whereas upward fluctuations always increase congestion, downwards fluctuations cannot "anticipatorily save" on it. This, in the end, implies that queues of some positive length should be expected to persist even in the long-run.

Thus, let us maintain the assumption that $\rho < \rho_c$. Then, the arrivals and departures from each node i follow Poisson processes with rates equal to $\alpha_i = \rho\frac{\beta_i}{n-1}$ and unity, respectively. Denote by p_{im} the steady state probability of a queue of size m in node i (i.e. the probability that there is a load of m pending problems being faced by node i). The induced probability distribution $(p_{im})_{m=0}^{\infty}$ must satisfy:[8]

$$\begin{aligned} \alpha_i p_{i,m-1} + p_{i,m+1} &= (\alpha_i + 1)p_{im} \qquad\qquad (m = 1, 2, ...) \\ p_{i1} &= \alpha_i p_{i0} \end{aligned}$$

The left-hand side of the first equation is the mean flow rate into the state m. That is, it adds the transition rate from state $m-1$ to state m (the queue has $m-1$ elements and a new problem arrives) plus the rate from $m+1$ to m (the queue has $m+1$ elements and a problem is solved). There are no other possible transitions into state m, since the arrival or departure of two problems at the same time has probability zero in a continuous-time Poisson process. On the other hand, the right-hand side of the first equation represents the flow out from state m, i.e. it adds the rates at which a queue that has m problems receives one more, or solves one. In sum, therefore, the first equation only says that in a steady state the flow into any given state has to be equal to the flow out of that state. The second equation is just like the first one, except that it reflects the simple fact that a queue in state $m = 0$ cannot go to state $m = -1$, since a problem can only be tackled when it arises.

The solution to the system of equations above can be checked to be:

$$p_{im} = (1 - \alpha_i)\alpha_i^m, \ m = 0, 1, 2, \ldots$$

Therefore, the expectation for the length of the queue at node i in the steady

[8] See Allen (1990) for a good introduction to queueing theory.

state, which we denote by λ_i, is:

$$\lambda_i = \sum_{m=0}^{\infty} m(1-\alpha_i)\alpha_i^m = \frac{\alpha_i}{1-\alpha_i}.$$

Over the whole network, the total expected length of the queues, i.e. the expected size of what might be called the *stock of floating problems* is (using (4))

$$\lambda(\rho) = \sum_{i \in N} \lambda_i(\rho) = \sum_{i \in N} \frac{\rho \frac{\beta_i}{n-1}}{1-\rho \frac{\beta_i}{n-1}}. \tag{6}$$

This magnitude, in turn, has its mirror image in the time dimension, where it shows as the average delay, say $\Delta(\rho)$, involved in solving problems. By the so-called Little's Law,[9] it follows that

$$\Delta(\rho) = \frac{1}{n\rho}\lambda(\rho).$$

[9]Proofs for this Law can be found in Little (1961) and Stidham (1974). A simple proof, which we adapt from Bentley (2000) is the following. Define $X(T) = C(T)/T$, as the rate of problems solved up to a certain period T, where $C(T)$ is the number of problems solved up to that period. Let $Z(t)$ denote the stock of problems in the system at time $t \in [0, T]$. Let $W(T)$ be the area under $Z(t)$ from 0 to T, which represents the total aggregated waiting time over all problems in the system in that interval. The mean waiting time per problem solved is defined as $R(T) = W(T)/C(T)$. The mean number of problems in the system is the average height of $Z(t)$, which is $L(T) = W(T)/T$. Clearly, $L(T) = R(T)X(T)$. On the other hand, by definition, we have that $\lim_{T\to\infty} L(T) = \lambda$, and $\lim_{T\to\infty} R(T) = \Delta$. Since, in a steady state, the average number of exits from the system per unit of time must equal the number that enter the system, it follows that $\lim_{T\to\infty} X(T) = n\rho$. Thus, $\lambda = \Delta \cdot n\rho$, which is the desired conclusion.

Intuitively, this merely reflects an "accounting identity": on average, the stock of floating problems $\lambda(\rho)$ is to be viewed as the result of the mean delay $\Delta(\rho)$ displayed by each of the $n\rho$ problems arising in the network per unit of time.

3.3 Designing the network for optimal performance

Once we understand the dynamics of a given network, we can address the issue of what is the optimal network layout of an organization, given that it involves some pre-specified set of nodes and has a given number of links at its disposal.

First, we introduce some notation. Given any network Γ, denote by λ^Γ, ρ_c^Γ, β_i^Γ, the value that the variables λ, ρ_c, β_i take for this network. Now let $\mathcal{U}$ stand for the set of all networks that can be constructed with a certain number of nodes and links, and denote by λ^* the lower envelope of $\{\lambda^\Gamma\}_{\Gamma\in\mathcal{U}}$, i.e.

$$\lambda^*(\rho) \equiv \min_{\Gamma\in\mathcal{U}} \lambda^\Gamma(\rho)$$

with

$$\mathcal{N}^*(\rho) \equiv \arg\min_{\Gamma\in\mathcal{U}} \lambda^\Gamma(\rho).$$

Since

$$\lambda^\Gamma(\rho) = \sum_{i\in N} \frac{\rho\frac{\beta_i^\Gamma}{n-1}}{1-\rho\frac{\beta_i^\Gamma}{n-1}} \tag{7}$$

it obviously follows that

$$\begin{aligned} \lambda^*(0) &= 0 \\ \lim_{\rho\uparrow\rho_c^\Gamma} \lambda^*(\rho) &= \infty. \end{aligned}$$

For any $\rho < \bar{\rho}_c \equiv \max_{\Gamma\in\mathcal{U}} \rho_c^\Gamma$, the lower envelope $\lambda^*(\rho)$ defines the optimal performance (i.e. lowest stock of floating problems) displayed by an organization

which faces the demands (nodes) and limitations (links) embodied by $\mathcal{U}$. Correspondingly, $\mathcal{N}^*(\rho)$ specifies the optimal network architectures (in general not unique) that underlie such an optimal performance. Our aim here is to characterize the *topological properties* of the networks in $\mathcal{N}^*(\rho)$ for each $\rho < \bar{\rho}_c$. In particular, for any such network Γ (and their corresponding β_i^Γ), we shall focus on its *polarization* $\theta(\Gamma)$, which is defined as follows:

$$\theta(\Gamma) = \frac{\max_{i \in N} \beta_i^\Gamma - \left\langle \beta_i^\Gamma \right\rangle}{\left\langle \beta_i^\Gamma \right\rangle}$$

For the moment, let us maintain the tentative assumption that, for each $\rho < \bar{\rho}_c$, all networks associated to $\mathcal{N}^*(\rho)$ display the same polarization and denote this value $\theta^*(\rho)$.

It is intuitive that the following two properties should hold for an optimal network. First, for ρ low, congestion is not expected to be an issue. Thus, optimality should involve minimizing distance, which is achieved by a network with the highest polarization: a star (or star-like) network. That is, for low values of ρ, we would expect $\theta^*(\rho)$ to take the highest possible value. On the other hand, as ρ draws close to the maximum value given by $\bar{\rho}_c$, congestion must become the crucial factor, and optimality should involve a balanced (symmetric) network. That is, $\theta^*(\rho)$ would take the smallest possible value for such high ρ.

To cast the previous discussion in more formal terms, note that, for low ρ (i.e. as $\rho \downarrow 0$), the performance of a network Γ can be approximated as follows:

$$\lambda^\Gamma(\rho) = \sum_{i \in N} \frac{\rho \frac{\beta_i^\Gamma}{n-1}}{1 - \rho \frac{\beta_i^\Gamma}{n-1}} \approx \frac{\rho}{n-1} \sum_{i \in N} \beta_i^\Gamma .$$

Therefore, for low ρ ("slightly above" zero), the task of finding the optimal networks in $\Gamma^*(\rho)$ involves singling out those networks Γ that minimize the aggregate betweenness $\sum_{i \in N} \beta_i^\Gamma$.[10] It is easy to verify that this minimization is attained

[10] If we define logarithmic distance as the average number of nodes that a problem has to travel in order to reach its destination, aggregate betweenness is equivalent to algorithmic distance. To see this, note that every time a problem goes from one node to another, it

by a star-like network where the polarization is maximal,[11] as indeed suggested above.

Instead, for high ρ (i.e. as $\rho \uparrow \rho_c^\Gamma$), the stock of floating problems (which rises unboundedly with ρ) is of the following order:[12]

$$\lambda^\Gamma(\rho) \sim \mathcal{O}\left(\max_{i\in N} \frac{1}{1-\rho\frac{\beta_i^\Gamma}{n-1}}\right) = \mathcal{O}\left(\frac{1}{1-\frac{\rho}{n-1}\max_{i\in N}\beta_i^\Gamma}\right).$$

This implies that, for high ρ ("slightly below" $\bar{\rho}_c$), optimal performance is achieved by networks Γ with a minimum value for $\max_i \beta_i^\Gamma$. Thus, as suggested in our discussion, the optimal network in this case is to be an homogenous one, where the maximum betweenness is minimized and thus polarization is minimal.

As ρ rises from very low levels to values close to $\bar{\rho}_c$, it is natural to conjecture that the optimal level of polarization $\theta^*(\rho)$ should vary in a monotonic (non-increasing) fashion. To check the validity of this conjecture, it is useful to turn our attention to the form of the objective function $\lambda^\Gamma(\rho)$ which is minimized over $\Gamma \in \mathcal{U}$ (cf. (7)). A first useful observation in this respect is that the dependence of this function on Γ is solely channeled through the corresponding vector of induced betweenness β^Γ. Thus, for each $\rho < \bar{\rho}_c$, we may equivalently reformulate

increases both its algorithmic distance by 1 unit and the betweenness of the receiving node by 1 unit.

[11] To see this simply note the following. First, the topological betweenness is never higher than the algorithmic betweenness – recall Footnote 5. Second, the topological betweenness is minimized at a star network, where the average (topological) distance is minimized. Third, at a star network, both notions of betweenness (topological and algorithmic) coincide.

[12] We say that $f(\rho) \sim \mathcal{O}(g(\rho))$ if $0 < \lim_{\rho\to\rho_c^\Gamma} \frac{f(\rho)}{g(\rho)} < \infty$.

the optimization problem underlying $\theta^*(\rho)$ as follows:

$$\min_{\beta \in \mathcal{B}} \lambda^{\beta}(\rho) \equiv \sum_{i \in N} \frac{\rho \frac{\beta_i}{n-1}}{1 - \rho \frac{\beta_i}{n-1}}.$$

Then, to proceed formally, we would need a sufficiently detailed characterization of the range of feasible betweenness vectors

$$\mathcal{B} \equiv \{\beta = (\beta_i)_{i \in N} \in \mathbb{R}^n_+ : \beta = \beta^{\Gamma} \text{ for some } \Gamma \in \mathcal{U}\}$$

that can be spanned by the set of admissible networks $\mathcal{U}$. This, unfortunately, seems an especially difficult task, given the complex combinatorial considerations involved. We may hope, however, to shed some light on the problem if we rely on the following two simple features of the situation. First, we note that $\lambda^{\beta}(\rho)$ is an increasing and convex function on $\mathbb{R}^n_+$ whose curvature increases with ρ. Thus, in particular, its level curves $\{\beta : \lambda^{\beta}(\rho) = K\}$ pass from being linear when $\rho = 0$ to displaying a "right-angle kink" at points of uniform betweenness (i.e. in the bisectrix) as $\rho \to \bar{\rho}_c$ (cf. Figure 1).

A second important observation derives from an already explained fact: the sum of betweenness is minimized over the set $\mathcal{B}$ at star-like configurations. To help formalize the implications of this fact, suppose that "perfect-star" networks with just one node i at the hub and all other nodes $j \neq i$ as symmetric pure spokes are admissible configurations in the set $\mathcal{U}$. Then, if we denote such star networks by $\widehat{\Gamma}^i$, it follows that the set $\mathcal{B}$ must lie above the following hyperplane in $\mathbb{R}^n$:

$$H \equiv \{\beta = (\beta_i)_{i \in N} \in \mathbb{R}^n : \sum_{i=1}^{n} \beta_i = \beta^{\widehat{\Gamma}^i} \text{ for any } i \in N\}.$$

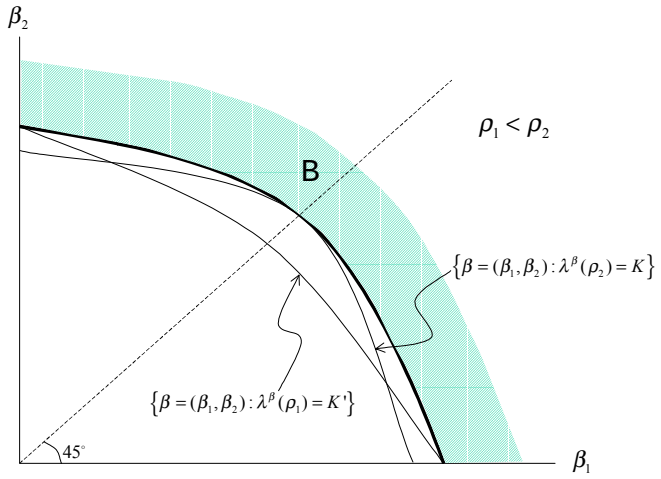

Figure 1: Optimal betweenness profile $\beta^*(\rho)$ as ρ passes from a relatively low $\rho = \rho_1$ to a higher $\rho = \rho_2$. For the lower ρ, the level curves display less marked curvature and the optimal profile occurs at the two extreme betweenness points where the corresponding level curve and the lower frontier of $\mathcal{B}$ meet in each of the two axes. For the higher ρ, the optimal profile lies at the tangency between the corresponding level curve and the lower frontier of $\mathcal{B}$ that lies in the bisectrix of the positive orthant.

Let us now make the plausible assumption[13] that the lower frontier of $\mathcal{B}$, i.e.

$$\partial\mathcal{B} \equiv \{\beta = (\beta_i)_{i\in N} : [\beta' \in \mathcal{B},\, \beta'_i < \beta_i \text{ for some } i \in N] \Rightarrow \left[\beta'_j \geq \beta_j \text{ for some } j \in N\right]$$

does not change curvature throughout the space $\mathbb{R}^n_+$. (An illustration of the situation is again provided in Figure 1 for the bidimensional case.) Then, combining the above considerations, it readily follows that, as suggested above, the polarization $\theta^*(\rho)$ associated to the optimal network depends on ρ in a weakly monotonic (non-decreasing) fashion. But the analysis can go much farther than this anticipated dependence and arrive at the following startling conclusion. As the problem rate ρ rises (and the "bending" of the level curves becomes progressively more acute) there is a *threshold transition* from the case where the optimal network displays a polarized betweenness (i.e. it is star-like) to a situation where the betweenness vector is essentially symmetric (and the network is basically homogenous). Thus, what this analysis suggests is that, as ρ changes, there is a qualitative "discontinuous" change in the optimal network that basically reduces the range of optimal configurations to two extreme cases: a fully centralized and a fully decentralized network.

We have checked the conclusions derived from this analysis (in particular, the validity of our simplifying assumptions) by exploring matters numerically in a variety of computationally amenable contexts. The results are shown in Figure 2 for the leading scenario described in Subsection 3.1 and a range of different possible specifications of $\mathcal{U}$ (i.e. different number of nodes and possible links).

Figure 2 plots the value of $\theta^*(\rho)$ as a function of ρ, for organizations that differ in the number of links (64, 96, 128, 160). The organizational size is kept constant at $N = 32$. The value of $\theta^*(\rho)$ was obtained through algorithmic search over the set of admissible networks.[14] In all cases, we observe that the degree of polarization associated to the optimal architecture depends on ρ as predicted by theoretical model, i.e. it is non-increasing in ρ and displays an abrupt change

[13] We ignore at present how restrictive this assumption really is, although we conjecture that it may be a sufficiently good approximation of the situation when the problem at hand is large enough, i.e. there are sufficiently many nodes and the considerations pertaining to "node indvisibility" are of second-order importance. As reported below, the simulations conducted in different (small) contexts provide indirect evidence in support of this conjecture.

[14] Let us explain the method used to perform the numerical search for the optimal network. We use generalized simulated annealing, as described in Penna (1995) and Tsallis and Stariolo

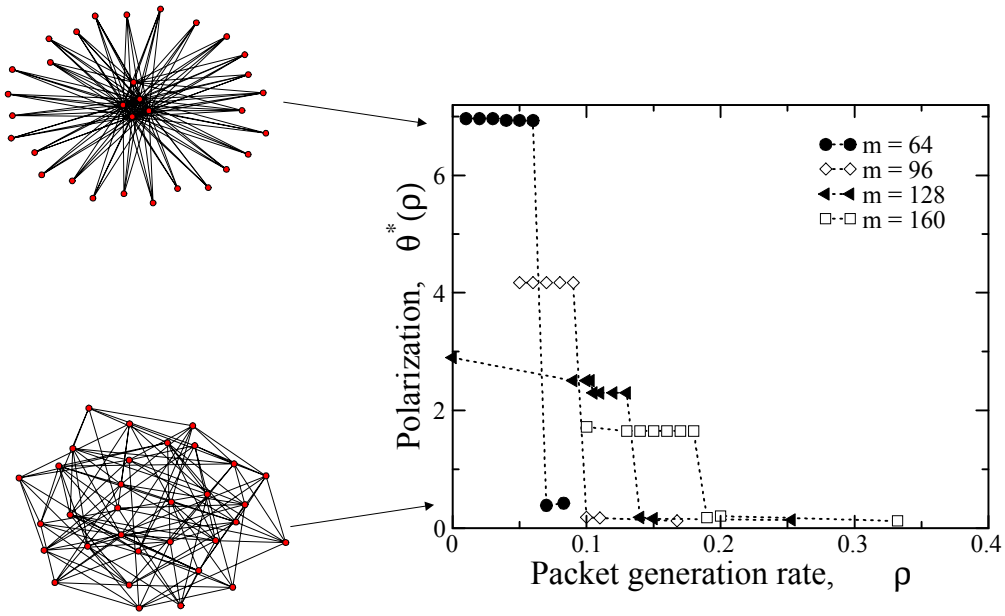

Figure 2: Polarization of the optimal structure as a function of ρ, for networks of size $n = 32$ and different number of links $m = 64, 96, 128, 160$. The star-like configuration (top left) is optimal for low ρ , while an homogeneous configuration (bottom left) is optimal for high ρ .

between the two extreme topologies – i.e. star-like and homogenous – as ρ varies.

(1994). Starting from a given initial network configuration, random rewiring of individual links are performed. The cost $\lambda^{\Gamma}(\rho)$ is then evaluated. The change is accepted with a certain probability that depends on a computational temperature. This temperature is decreased with time so that the system tends to explore regions of the configuration state with lower and lower costs.

Regarding the cooling, at a given temperature, each node of the network is allowed to try a rewiring. Then the temperature is decreased by 1%, and the process is repeated until a minimum temperature is reached or, alternatively, the system has remained unchanged after a significantly large amount of rewiring trials.

Different sets of initial conditions are explored: for a given value of ρ, the optimization

Moreover, throughout the whole range of ρ, *only these two topologies* ever qualify as optimal.

4 Related literature

In the last few years there has been a booming interdisciplinary interest in the study of networks. Social scientists have been working steadily on this topic, but also physicists interested in the dynamics of complex systems, or biochemists studying autocatalytic networks and the origin of life. This vast line of research has been motivated by the belief that social, physical, or biological models that ignore the topological structure of interaction are often unable to give account of many interesting phenomena. The increasing importance of the world wide web for scientific, governmental and commercial purposes is another powerful source of interest in this topic.

Our paper belongs most directly to the literature on the economics of organizations. In a sense, our analysis reflects the same informational considerations that have long lied at the core of the controversies on the merits and drawbacks of economic (de)centralization.[15] However, rather than highlighting how the richness of information or the cost of communication bears on the problem, our analysis displays a somewhat different focus. We stress that limitations on the ability to process a large amount of information simultaneously raises the threat of organizational collapse or at least long delays in the organization tackling the required tasks.

There is a recent strand of the economic literature that is motivated by similar concerns and also identifies organizations with networks whose objective is to *process information*. The paper initiating this line of research was Radner (1992), then followed, among others, by Bolton and Dewatripoint (1994), and van Zandt (1999). Their work mostly abstracts from search issues. The information that flows in an organization is such that any of its members can process it. Typically, there are advantages in terms of processing time if different bits of the same problem are processed in parallel. But, in this case, the different bits must be combined in order to obtain the final output, and the required communication

process is started from random initial configurations and also from networks that turned out to be optimal at similar values of ρ. Of all the realizations, only the network with a smallest cost is considered as optimal.

[15] This debate, for example, is nicely epitomized by the well-known work of Lange (1936, 1937) and Hayek (1940). The central issues raised by these authors were later formulated and addressed formally by the Theory of Mechanisms, as initiated by Hurwicz (1960). See van Zandt (1999) for a good survey on this topic.

brings about a coordination problem. The main trade-off here is the one between parallelization and coordination costs. The organization consists, thus, of a rather mechanical process of *combining* disperse information. Sah and Stiglitz (1986) and Visser (2000) also study an analogous design problem, their main focus being on the contrast between the performance of a hierarchic and a poliarchic organization.

Closer in spirit to our work is Garicano (2000). In his model, each individual specializes in solving a certain type of problems. If she cannot solve a problem that reaches her, there is another person to whom she must deliver that problem. The task of the organization designer is twofold. First, she must assign knowledge sets to each individual in the organization. Then, she must design the routes through which unsolved problems must travel. Both knowledge acquisition and communication are costly. There is, then, a fundamental trade-off between acquiring knowledge and communicating it. The solution to this trade-off is to organize workers along a hierarchy. All problems are first given to the workers lowest in the hierarchy, who have the knowledge about the most ordinary problems. Those relatively uncommon problems that they cannot solve are then transferred to individuals in the next higher level, and so on.[16]

Despite the similarity in spirit, there is a crucial difference between Garicano's (2000) model and ours. We assume that knowledge acquisition cannot be controlled or designed and thus the organization planner must take the knowledge sets of workers as given. This, in turn, creates a congestion problem in our set-up which does not appear in his context. Since the planner in Garicano (2000) has control about what every worker knows, the organization can be designed so that bottlenecks are avoided. We feel that our model is relevant for firms in which endowments of knowledge are not easy to replicate in a standardized fashion. Even if a university wanted, it would be hard to find two solvers of Fermat's last conjecture for every ten solvers of standard elliptic partial differential equations. We conjecture that the high-level knowledge-based organizations we used to mo-

[16]Beggs (2001) introduces a model that is close (and produces similar conclusions) to Garicano (2000), with two important differences. From the conceptual point of view, the differences between workers in Beggs (2001) arises because of different ability (processing power) between individuals, rather than because of specialization, as in Garicano (2000). From the technical point of view, Beggs (2001) uses an explicitly stochastic model, and the techniques come mainly from queuing theory. The difference between individuals in our model occurs because of specialization, so in that sense we are closer to Garicano (2000). In the technical respect, however, we are closer to Beggs (2001), which also makes a imporatnt use of queueing theory.

tivate our paper present characteristics that make them look more like those in our model.

A more technical literature has focused in the problem of search in complex networks. Watts and Strogatz (1998) pioneered the recent surge of interest in what has been called *small-worlds* (see also Watts 2000, and Newman, Moore and Watts 2000). This term refers to regular lattices where nodes have many local links (links that connect nodes to neighbors in an underlying topological sense) and a few long-range links. This kind of networks have the characteristic that the average distance between two randomly chosen nodes is relatively low. This is so despite the fact that most connections are purely local. The small-worlds literature abstracts from search problems (and also congestion), since distance here means minimal graph distance and thus implicitly presumes global knowledge of the network. Albert and Barabási (2002) survey the findings in the area.

Kleinberg (1999, 2000), on the other hand, does address search issues in the context of complex networks. In his model, problems have to travel through a network looking for its (known) destination. The search is helped by knowledge of the underlying "geographic structure" (and the links of each node). This structure may be very effective in guiding search within a small-world type network. In contrast, it is not useful in a random network (i.e. one whose links are completely random), despite the fact that average distance is actually smaller. Kleinberg's model helps to explain the speed and effectiveness of search in some large complex networks (e.g. the huge world-wide web). It abstracts, however, from the congestion issues that are our main interest here and that, undoubtedly, also represent a key consideration in many real-world contexts. Arenas, Díaz-Guilera and Guimerà (2001) address problems similar to those considered here and study, in particular, the trade-off between congestion and distance. They restrict, however, to a limited range of possible organizational forms, namely hierarchies, which face no genuine issue of search. In a hierarchy, all problems (which are aware of their destination) know the (fixed) route they have to travel.

5 Summary and extensions

We have proposed an abstract model of a problem solving organization which (a) operates through local communication, (b) is forced to search restricted by local information (c) is subject to the effects of congestion. For this model, we provide an analytical characterization of the threshold of collapse and the stock of floating problems (or average delay) below that threshold. We then build upon this characterization to start addressing a design problem, namely to find the network which optimizes performance for any given problem arrival rate.

A number of extensions could be explored. An interesting one concerns studying the effect of a larger "information radius" on the performance of the organization. That is, when designing the optimal organization, we assumed that individuals only use information about their direct neighbors to route a problem.

We are currently undertaking research to relax this assumption. Individuals may use the knowledge of their neighbors'connections (or even of individuals with higher order degrees of separation). First, concerning the issues of congestion and delay, it is easy to see that the analytical approach used here to characterize the congestion threshold and the average delay may be applied unchanged for any information radius (remember we only started use the assumption of first neighbors knowledge for the design problem). Turning then to the issue of organizational design, preliminary numerical results suggest that, as one would expect, the optimal network becomes less polarized as the information radius expands. This is intuitive since, as the information of nodes becomes less local, the informational advantages of a polarized network should correspondingly decrease.

Many other extensions could be easy to handle in our framework. For example, the problems could be sent with higher (or even lower) probability to nodes with a larger number of connections. Also, the rate at which problems originate at one node could depend on the node where they can be solved, which may create local "communities" of problem-solvers.

References

R. Albert and A.L. Barabási (2002), "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* 74:47-97.

O. Allen (1990), *Probability, Statistics and Queueing Theory with Computer Science Application* New York: Academic Press.

A.W. Beggs (2001), "Queues and Hierarchies," *Review of Economic Studies* 68:297-322.

J. Bentley (2000), *Programming Pearls*, Boston: Addison-Wesley.

P. Bolton and M. Dewatripont (1994), "The Firm as a Communication Network," *Quarterly Journal of Economics* 109:809-839.

A. Arenas, A. Díaz-Guilera, R. Guimerá (2001), "Communication in Networks

with Hierarchical Branching," *Physical Review Letters* 86:3196-3199.

L. Garicano (2000), "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy* 108:874-904.

F.A. Hayek (1940), "Socialist Calculation: The Competitive Solution," *Economica* 7: 125-49.

L. Hurwicz (1960), "Optimality and Informational Efficiency in Resource Allocation Processes," in *Mathematical Models in the Social Sciences,*, ed. K.J. Arrow and L. Hurwicz, Cambridge: Cambridge University Press.

J. Kleinberg (1999), "The Small World Phenomenon: An Algorithmic Perspective," Cornell Computer Science Technical Report 99-1776.

J. Kleinberg (2000), "Navigation in a Small-World," *Nature* 406:845.

O. Lange (1936), "On the Economic Theory of Socialism: Part One," *Review of Economic Studies* 4: 53-71.

O. Lange (1937), "On the Economic Theory of Socialism: Part Two," *Review of Economic Studies* 4: 123-42.

J.D.C. Little (1961), "A Proof of the Queueing formula: $L = \lambda W$," *Operations Research* 9:383-387.

S. Stidham Jr. (1974), "A Last Word on $L = \lambda W$," *Operations Research* 22:417-421.

M.E.J. Newman, C. Moore, and D.J. Watts (2000), "Mean-Field Solution of the Small-World Network Model," *Physical Review Letters* 84:3201-3204.

T.J.P. Penna (1995), "The Travelling Salesman Problem and Tsallis Statistics," *Physical Review E* 51:R1-R4

R. Radner (1993), "The Organization of Decentralized Information Processing," 61:1109-1146.

R.K. Sah and J.E. Stiglitz (1986), "The Architecture of Economic Systems: Hierarchies and Polyarchies," *American Economic Review* 76:716-727.

C. Tsallis and D.A. Stariolo (1994), "Optimization by Simulated Annealing: Recent Progress," in *Annual Review of Computational Physics,* II:343, ed. D. Stauffer, Singapore: World Scientific.

B. Visser (2000), "Organizational Communication Structure and Performance," *Journal of Economic Behavior and Organization* 42:231-252.

D.J. Watts and S.H. Strogatz (1998), "Collective Dynamics of 'Small-World' Networks," *Nature* 393:440-442.

D.J. Watts (1999), *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton: Princeton University Press.

T. van Zandt (1999), "Decentralized Information Processing in the Theory of Organizations," in *Contemporary Economic Development Reviewed, vol. 4:*

The Enterprise and its Environment, ed. M. Sertel, London: MacMillan.

T. van Zandt (1999), "Real-Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents," *Review of Economic Studies* 66:633-658.