Collaboration, Motivation, and the Size of Organizations

Bernardo A. Huberman

Dynamics of Computation Group
Xerox Palo Alto Research Center
Palo Alto, CA 94304
E-mail: huberman@parc.xerox.com

Christoph H. Loch*

INSEAD
Boulevard de Constance
77305 Fontainebleau, France
E-mail: loch@insead.fr

Short Title: **Collaboration, Motivation, and Organizational Size**

Corresponding author:

Christoph H. Loch
INSEAD
Boulevard de Constance
77305 Fontainebleau, France

Phone: 011 33 1 6072 4326
E-mail: Loch@insead.fr

* The author was supported on this project by the INSEAD Research and Development fund.

Collaboration, Motivation, and the Size of Organizations

Bernardo A. Huberman

Dynamics of Computation Group
Xerox Palo Alto Research Center
Palo Alto, CA 94304
E-mail: huberman@parc.xerox.com

Christoph H. Loch*

INSEAD
Boulevard de Constance
77305 Fontainebleau, France
E-mail: loch@insead.fr

Abstract

We present an analytical model of an organization that offers operational drivers of limits on team size. The model trades off benefits from collaborative problem solving against the disadvantages of diminishing motivation when groups get large. Collaboration is represented as parallel employee activity combined with frequent sharing of partial information. The resulting performance increase over team size is between linear and quadratic. Motivation is modeled by each team member periodically setting an effort level either to contribute to the best of one's ability or to "cruise" at the minimum level not recognizable as shirking. Each individual decision is limited by bounded rationality based on team rewards, the time horizon of team interaction, and individual expectations about colleagues' behavior. The decision collapses to a simple "barrier rule": Work hard when a certain "barrier percentage" of team members work hard, and otherwise shirk. When the performance increase from collaboration is linear, this barrier percentage increases with team size and eventually grows beyond 100%, preventing cooperation for large teams. When performance increases quadratically in team size, however, collaborative problem solving benefits balance motivational effects, resulting in the possibility of sustained cooperation up to an externally given limit (such as coordination costs). Thus, effective problem solving methods are of double value, improving direct productivity *and* mitigating the social dilemma from team production. A manager should enlarge his organization up to the minimum of the limit set by the cooperation barrier and the exogenous performance limit.

KEYWORDS: Cooperative problem solving, cooperative processes, social dilemma, team production, expectations, bounded rationality, organizational size.

* The author was supported on this project by the INSEAD Research and Development fund.

1. Introduction

Our knowledge about the optimal size of organizations and organizational units is rudimentary and incomplete. This is exemplified, first, by the fact that corporations vary vastly in size. The global food company Nestle, for example, has a rule of thumb that development units should not be larger than 200 employees. Automobile factories, on the other hand, commonly combine more than 10,000 employees in a tightly managed unit. If organizational principles are behind such large differences, one would like to know what they are.

Second, there seems to be a recent trend towards smaller firms, triggered by the rise of new firms in such industries as electronics, pharmaceuticals, and biotechnology. It is often argued that this trend is caused by new information technologies [8]. If that is the case, it is of interest to find how these technologies change the usual tradeoffs involved in determining organizational size (see Brynjolffson and Mendelson [7] for a qualitative discussion and Barron [4] for a model of a hoerarchical organization). And more important for the manager leading an organizational unit is understanding what changes can enable a desired increase or decrease in unit size.

This article makes two contributions towards answering these questions. First, we introduce a previously unstudied phenomenon, namely collaborative problem solving. Second, we examine the tradeoff between the benefits of group size on collaborative problem solving and the negative impact of larger groups on work motivation (e.g., shirking). This analysis is relevant not only to the aggregate level of firms as a whole, but also to the *operational level of managing an organizational unit*. Our model derives operational drivers of group performance, such as team size, problem complexity, cooperation intensity and quality, and group rewards. It also allows parameter adjustment by managers interested in improving organizational performance, adjustments that depend on the complexity of the problems to be solved.

We accomplish our analysis by viewing an organizational unit as a *problem solving unit* under common supervision and with a quantitative measure of group performance. The problems may be simple or complex and may represent market analyses, development activities, or production of any kind. It is well known that group collaboration improves problem solving capabilities. For example, productivity normally increases with the number of ideas an engineer develops in discussions with peers, up to a limit determined by the fact that efforts of further communication outweigh the benefits (see, for example, Allen [2]). We quantify this observation in an explicit model of collaborative problem solving, which leads to better performance with increasing team size.

On the other hand, it is also known that a single employee's influence on team performance diminishes with increasing team size, causing a temptation to shirk in a tradeoff between more leisure[1] and marginal performance benefit. For example, employees may perform only enough to avoid a negative impression on the part of the group manager, instead of actively trying to improve output as much as possible. In most organizations, supervision cannot possibly prevent employees from such behavior. Thus, shirking limits the effective size of a problem solving organization. In deciding about shirking or full cooperation, employees do *not* carry out a full game-theoretic analysis of all possible future actions by themselves and their peers. They most often exhibit bounded rationality, acting upon limited information and certain expectations about behavioral norms in the group. We explicitly include bounded rationality into the model by incorporating a reasonable form of expectations.

The analysis of the model yields two organizational design principles: one, in many organizations the individual's decision to cooperate collapses to a simple "barrier rule": work hard when overall performance implies that a certain "barrier percentage" of team members are working hard, and otherwise shirk.

Two, the influence of team size on this barrier percentage depends on the extent of benefits from collaboration: The barrier increases with team size and eventually grows to 100% if performance improvement is only linear in the number of cooperating team members. With a quadratic performance increase, however, the increased benefits that result from collaboration balance the temptation to shirk, with the barrier percentage approaching a fixed limit for large team sizes. This implies that cooperation is sustainable in large organizational units, provided the problem solving processes used are powerful enough to ensure sufficient performance increases. In practice, team performance cannot increase indefinitely with team size, due to communication costs or theoretical performance limits. The normative conclusion of the model is thus: concentrate effort on problem solving methods, since their effectiveness directly improves productivity *and* mitigates the social dilemma inherent in team production. The manager should enlarge the team until either the cooperation barrier or external performance limits negate team cooperation.

---

[1] Here, the word leisure is used in the same way as in Alchian and Demsetz [1]: Leisure stands for "all nonpecuniary income for simplicity of exposition". Examples are actual free time, but also the use of company resources for private purposes (e.g., telephone calls).

Three, we demonstrate how team size is influenced by operational parameters in the model, such as the team's time horizon, bonus size, problem solving methods, and communication quality and quantity.

This paper is organized as follows. Section 2 gives a brief overview of two relevant literatures: the theory of the firm within the field of industrial organization, and that of distributed intelligence and cooperative problem solving, developing the methodology on which the analysis in this paper rests. Section 3 presents the basic model of collaborative problem solving and of team motivation with bounded rationality. Section 4 derives the associated organizational size for a linear increase of collaborative benefits, and Section 5 examines the effect of a quadratic performance increase on team size. In Section 6, we discuss open questions and give an outlook of further research.

2. Overview of Past Work

Our approach to the question of organizational size draws on two literatures: collaborative problem solving and social dilemmas in groups. Our representation of collaborative problem solving is based on recent work in the field of Artificial Intelligence. This work quantifies the widely shared observation that a group of coordinating "agents"[2] engaged in problem solving can solve a task faster than either a single agent or the same group of agents working in isolation from each other. The theory is developed in Huberman [14] and experimentally tested in computer experiments through simulations by Clearwater *et al.* [10].

The theory considers systems of team members collectively engaged in solving a complex task. The solution process is represented by heuristically searching a problem space, represented as a search tree. This is a very general problem representation, which is well known in Operations Research and applicable to many unstructured problems in organizations. Team members are assumed to work on the problem in parallel and to be diverse in their solution methods, meaning that they use differing heuristics. During the search process, team members exchange "hints," i.e., information about partial solutions and areas of the search space that can be excluded from further consideration. The search is successful for the team as a whole as soon as the first member has found a solution to the problem.

---

[2] The artificial intelligence literature uses the term "agents" in the sense of intelligent actors in an interactive system. The Principal-Agent literature in Industrial Organization refers to agents as employees with an own objective separate from that of the firm's owner. We use the term with both meanings, and also interchangeably with "team members" or "employees". In referring to an agent, we use "her" and "his" as synonyms.

It was found that such highly cooperative systems, when sufficiently large, display a universal distribution of individual team member performance, namely a log normal, which is largely independent of the detailed nature of the heuristics employed or the particular problem being tackled. This prediction was tested in computer simulations and is supported by a number of empirical studies on individual productivity and problem solving in the scientific community [6].

At the level of the team as a whole, the theory predicts an extended high performance tail of the performance distribution. In other words, the sharing of hints can greatly benefit a few team members and allow them to find a solution very fast. This effect becomes more pronounced the more hints are exchanged and the higher their quality, and the more diverse the team members. The relevant result for our present paper is that performance of the team as a whole (speed to solution of the task) can increase faster than linearly with team size. In some computer simulations, quadratic improvements have been observed.

This benefit of team size must be traded off against the negative effects that a large group size has on individual motivation. The collective contribution of all members of a team or organizational unit to overall performance represents a situation known as *social dilemma,* which has a long history in political science, sociology, and economics. The model in the present paper is based on Glance and Huberman [12]. In this specific social dilemma, the team members of an organizational unit jointly produce a measurable team output in the presence of weak central authority. The team output can be thought of as the number of problems solved per unit time or the quality of the solutions found; it can also represent the number of designs produced in engineering, or the number of process problems fixed in manufacturing.

Within this context, team performance represents a common good, because it is collaboratively produced and no member's contribution is separable and recognizable except through monitoring. The performance measure is associated with some performance-related part of salary, such as a bonus. Usually, management authority is "weak" (imperfect), in the sense that it is impossible to monitor each team member sufficiently to know whether he or she is contributing to the best of her abilities or only "cruising" at a comfortable effort level.

The social dilemma we have described can be represented using game theory. Each individual has two choices: either to *cooperate,* which means contribute to the best of her abilities, or to *shirk* and free ride on the work of others, i.e., put in only the minimum amount of effort to avoid "being caught." All individual team members share equally in the performance measure regardless of their actions. Each person that cooperates increases the team performance level by a certain amount, receiving the associated bonus increase in return. Whenever the cost of cooperating is greater than the marginal benefit, the individual defects. Since each individual faces the same choice, all defect and the team performance level stays low, at the minimum level. In the language of economics, the individual's rational strategy of weighing costs against benefits results in an inferior outcome, and no common good is produced.

However, the logic behind the decision to cooperate or not changes when the interaction is ongoing since future performance benefits will join present ones in influencing the rational individual's decision [13]. In most cases, employees do not base their decisions on a full game theoretic analysis of the situation. Bounded rationality limits them to taking into account limited information about current behavior, plus individual expectations concerning the future behavior of their colleagues. Two factors influence the individual's expectations: first, how long the interaction is expected to last. If the game is likely to end soon, then, rationally, future expected returns should be discounted heavily with respect to known immediate returns. On the other hand, if the interaction is likely to continue for a long time, then members may be wise to discount the future only slightly and make choices that maximize their returns in the long run. The second factor influencing expectations is the way in which a member believes her choices to influence the decisions others make. For example, if a team member believes that her actions influence contributors more than shirkers, then cooperation may be more self-enforcing than when individuals feel their own influence being diluted and opportunistically defect in the face of broad cooperation.

The analysis of this social dilemma leads to a critical group size beyond which cooperation is no longer sustainable. In large groups, the marginalized individual's temptation to free ride becomes too great, and shirking will be the usual answer. Thus cooperation in large groups with short expectations is hard to secure voluntarily.

The model we present in this paper combines explicit representations of cooperative problem solving and of shirking under imperfect monitoring, as well as the tradeoff between the two. This approach allows to address many operational questions of organizational design.

The question of optimal firm size is addressed by a literature in the field of Industrial Organization in economics. This literature is briefly reviewed below as it is closely related to the topic of this paper, even though it focuses on a different level of aggregation.

A classic argument for large organizations has been the existence of economies of scale in operations - procurement, production, and sales. This argument, however, does not hold for many processes beyond a certain size; for example, one complete fabrication line in semiconductor manufacturing. Moreover, it does not explain why any integration into organizations beyond the pure scale sensitive process is observed; for example, why procurement and production are not usually performed in different organizations and coordinated via market relationships.

Industrial Organization theorists give two answers to this question (for an overview, see Martin [17], p. 210 ff., and Milgrom and Roberts [18]). The first comes from transaction cost economics, which essentially states that relationships with high asset specificity and uncertainty are more efficiently governed through common ownership than through market contracts. If one party in a contract makes a specialized investment (such as a special machine) and the outcome of the investment is uncertain (perhaps because of unforeseeable market evolution), then this party wants to be protected from opportunistic behavior by the other party (e.g., letting the party committed to the specialized asset be responsible for all losses in case of unfavorable market development). As it is often impossible to include all contingencies in a contract, integration into one organization is one possible solution in such cases (see Williamson [21]).

The second answer relates to "team production" (Alchian and Demsetz [1]). It is defined as production using several employees (or resources), in which the output is a nonseparable sum of the efforts of the employees (and hopefully larger than the sum of separate efforts by the employees), i.e., each team member's individual contribution is not identifiable from the output. Nonseparability of the output results in the need for monitoring, because a team member has an incentive to shirk and gain leisure while imposing most of the lost output on the other team members. Since a specialized monitor on the team may himself have an incentive to shirk, the solution in neo-classic economic theory is to give one person the title to the net earnings of the team, net of all payments to the other team members. Thus, the monitor ("owner of residual rights") will ensure maximum team productivity without shirking.

Often, monitoring will not be completely possible as the cost of perfect observation of employees' behavior is prohibitively high, not to mention its demoralizing effects. In this case (the principal - agent problem), the owner (principal) will have to give the workers (agents) a performance incentive in order to prevent shirking (Mirrlees [19]). In this model, the supervision span (the number of agents controlled) by one principal is limited, because an increase in span causes the monitoring effort to outweigh the benefit from additional agents. Larger organizations can be built, however, via the introduction of hierarchies. The number of hierarchies itself is limited by the

"Law of Diminishing Control": Williamson [20] shows in a model with fixed span of control per hierarchical level that any positive uniform "control loss" between hierarchies (defined as a percentage of the supervisor's intentions not satisfied by the subordinate) limits the maximum size of the organization. The reason is that the control loss will make the cost of an additional management layer, if too many management layers are added, outweigh the contribution of the additional workers.[3]

Barron [4] develops a refined model of a hierarchical organization, with parameters for communication and control loss and hypotheses how these parameters are influenced by IT. We model communication and collaboration explicitly for one organizational unit, where Barron includes exogenous parameters and instead focuses on the corporate hierarchy as a whole. Thus, our model is more operational in character.

Industrial Organization is interested in the reasons why firms exist as players in industries. Consequently, transaction cost theory looks at the border between the firm and the market. It does not address the internal structure of firms into organizational units, and it does not give any hints about the possible size of organizations. The concepts of team production and agency, on the other hand, are much closer to the spirit of the present paper. However, they describe phenomena on the aggregate level of the firm as a whole, i.e., they do not fully spell out what team production and control loss mean operationally. This style of analysis is useful for questions in industrial organization, but less so for deriving managerial guidelines on the shaping of organizational units and their management.

3. A Model of a Problem-Solving Organization With Imperfect Supervision

Our model has two principal ingredients -- the benefits of cooperation in problem solving between team members, and the temptation to shirk caused by the collective nature of team output, which prevents managers or peers from observing each individual's contribution exactly. Before analyzing the tradeoff between the two, we will describe them in turn.

3.1. Cooperative Problem Solving

[3] Williamson's analysis is disputed by other authors [9], who argue that a control loss does not necessarily take place. A third group [16] offers an alternative concept of the cost of hierarchy, which they call "cost of coordination".

The simplest way in which cooperation improves the speed with which problems are solved in an organization is parallelism: Consider a task made up of many subtasks, each of which is assigned to an individual. If the problem contains T tasks allocated to n individuals, the whole process/problem can be done in T/n operation times, thus providing a speedup in performance that scales linearly with the number of individuals.

In most problems, however, not all tasks are independent of each other; this implies that the best performance will be determined by the longest sequential piece of the problem. For example, the sum of two numbers computed by different team members (agents) cannot be completed before both numbers have been computed. This implies that the best problem solving time to be expected is determined by the longest sequential path in the solution procedure. So if the problem contains P sequential operations, the total time will be $\frac{\text{time with 1 agent}}{\text{time with } n \text{ agents}} = \frac{T}{P+(T-P)/n}$. This equation is called *Amdahl's Law*, which implies that the total speedup from parallel processing cannot exceed P / T. Thus, the linear speedup predicted above cannot be achieved.

In collaborative problem solving, however, there is an added advantage stemming from the fact that individuals can help each other by exchanging information in the form of hints. Consider an organization of individuals with diverse characteristics trying to solve a constraint satisfaction problem. This is a problem in which values must be assigned to a set of variables so that a number of conditions (the constraints) are satisfied in order for the assignment to be a solution. Problems of this kind are very common in practice; for example, packaging and interior design in car development can be formulated as a constrained satisfaction problem.[4] The two design aspects, interior and exterior dimensions, are strongly constrained by product specifications and by each other; thus, no "optimum" exists.

In most cases of interest, no direct solution method is known, and one must resort to searching through a large number of possible assignments to find those that satisfy all the constraints. A *state* in the search is a set of assignments for all the variables, and a *partial state* has only some of the variables assigned. These search problems can be characterized by the total number of states in the search space, T, and the number of solutions, S. Let n be the number of agents involved in the search. In the complicated problems in question, we have $T >> S$ and $T >> n$. In typical practical problems, for example, the size of the search space T grows exponentially with the size of the problem.

---

[4] Note that no optimization is involved here; this fact corresponds to many practical situations where only a solution satifying minimum "goodness criteria" can be found. However, optimization can be incorporated into this theory in the form of search trees familiar from Operations Research.

The speed at which a team member can solve the problem depends on the initial conditions and on the particular sequence of actions as she moves through the search space. This sequence relies on the quality of the heuristics (the knowledge about the problem) the agent is using. The better the heuristics, the quicker she will be able to solve the problem. When many team members work on the same problem, this heuristic can include *hints* from colleagues suggesting where solutions are likely to be. It is important that the colleagues are *diverse* in the sense that they use different heuristics -- if they follow the same method, then information is duplicated without effect. The requirement of diversity is well supported by empirical studies, for example, in Angle's ([3], p. 145) demand for "communication among persons *unlike* one another, so that new, creative ideas are generated."

In computer experiments performed to test these ideas [10], the search problem example used was solving crypt arithmetic codes.[5] Hints consisted of assignments to a certain number of letters that the team members wrote to a central blackboard accessible to everybody. There was no prespecified limit to the size of the blackboard except that provided by the constraints of the problem. Other organizational structures can be modeled by using separate blackboards according to permissible information flows. Diversity among the team members was modeled simply by programming the agents to only examine hints of certain lengths.

In the experiments, members maintained partial number assignments that were consistent (added up correctly), periodically replaced these with hints of the right length, and generated random assignments for the non-consistent letters. Each agent performed a search step (solution update) according to a Poisson process of common rate $\lambda = 1$, in order to avoid synchronous "clockwork marching" by all team members.

In the context of the present paper, the problem is solved as soon as the first team member finds a solution.[6] The performance measure is thus the speed of finding the solution Π (the reciprocal of the time to find the solution). The speedup with an increase in team size can be characterized analytically [14], but the relevant result for the present paper comes from computer experiments. It shows a superlinear speedup as the number of diverse agents increases,[7]

---

[5] These problems require finding unique digit assignments to each letter of a word addition so that the numbers represented by the words add up correctly. An example is the sum DONALD + GERALD = ROBERT, which has a unique solution and a search space of size 10!.

[6] In very difficult problems, one must sometimes be content with a satisficing search, in which case the problem is solved when a solution of a certain quality is found.

[7] This is an *average* speedup, taken over many simulation runs in which the solution searches of team members unfold probabilistically.

$$\Pi = bn^{1+\varepsilon}, \tag{1}$$

where b is a proportionality factor, and the highest exponent $(1+\varepsilon)$ experimentally achieved is $(1+\varepsilon) = 2$, which corresponds to a *quadratic* performance increase with team size.

The superlinearity comes from the fact that team members with a given strategy can contribute hints to colleagues which they cannot use themselves. The colleagues can utilize the hints by using a different strategy. This leads to a non-linear increase of hints available. The increase in solution speed thus depends on the number of *diverse* team members using *different* solution strategies; a number of team members with identical strategies do not contribute to faster problem solving.

The parameters b and ε incorporate several operational characteristics of the organizational unit's processes; for example, quality of team members' solution strategies (how good are the individual heuristics, and how effectively do they incorporate hints?), quality of information exchange (do hints contain random noise or are they perfectly helpful?), organizational memory (how many hints can the blackboard hold?), and organizational structure (who communicates with whom? who can read and write on which blackboard?). We are not including these operational process characteristics in the present paper. We also ignore the cost of communication possibly outweighing collaborative benefits in large groups (this is discussed below). In the conclusion, we will give a brief outlook on how these issues can be incorporated in the theory.

3.2. The Dilemma of Team Cooperation

In our mathematical treatment of the collective action problem (based on Glance and Huberman [11]) we state the benefits and costs to the individual team member associated with the two actions of cooperation and defection (shirking). While no individual can directly observe the effort of another, each member observes instead the collective output and can deduce overall group participation. Each member sets individually his/her effort level.

Let k_i denote whether member i intends to cooperate ($k_i = 1$) or defect ($k_i = 0$). The number of members cooperating is $n_c = \Sigma\, k_j$. We can also introduce an amount of uncertainty into the relation between members' efforts and group performance. In this case, the perceived level of cooperation will differ from the actual attempted amount. For the purpose of this paper, we restrict attention to a deterministic, error-free world of team members' actions. For a description of the stochastic results of cooperation errors see [11].

According to Equation (1), we assume team performance Π to increase with the effective contributions of the members according to the relationship $\Pi = bn_c^{1+\varepsilon}$. We consider two benchmark cases: a linear performance increase ($\varepsilon = 0$), and a quadratic performance increase ($\varepsilon = 1$).

Each team member receives a fixed salary, w, plus a piece of the team's performance bonus of $\beta \Pi$, which is equally shared by all team members. Each contributing individual bears a personal cost, c. Then the utility at time t for member i is

$$U_i(t) = w + \frac{\beta}{n} bn_c(t)^{1+\varepsilon} - ck_i \tag{2}$$

Using her knowledge of the functional form of the utility function,[8] each individual can deduce the number of individuals effectively cooperating from the utility collected at time t by inverting Equation (2):

$$n_c(t) = \{\frac{n}{\beta b}[U_i(t) + ck_i - w]\}^{1/(1+\varepsilon)} \tag{3}$$

We also define $f_c(t)$ to denote the fraction, $n_c(t) / n$, of individuals effectively cooperating at time t. When all members contribute successfully, each receives net benefits of $w + \beta b n_c^{\varepsilon} - c$. However, the individual team member's marginal cooperation benefit is only $\beta b n_c^{\varepsilon}[1 - (1 - 1/n_c)^{1+\varepsilon}] - c$. This marginal benefit represents the utility difference between cooperation and shirking, assuming that all other team members continue to cooperate. Thus, team cooperation becomes a dilemma when

$$\beta bn^{\varepsilon} > c > \beta bn^{\varepsilon}[1 - (1 - \frac{1}{n_c})^{1+\varepsilon}]. \tag{4}$$

For our benchmark cases of linear and quadratic performance increase, this simplifies to

$$\beta b > c > \beta b / n. \qquad \text{(linear)} \tag{4a}$$

$$\beta bn > c > \beta bn(2 - \frac{1}{n}). \qquad \text{(quadratic)} \tag{4b}$$

---

[8] For a justification of the form of the individual utility function in the context of either divisible goods or pure public goods see Bendor and Mukherjee [5].

In the case of equation (4), the dominant strategy in the one-shot game is to defect, even though team performance, and thus everybody's bonus, is maximized when everyone cooperates. The reason is that the additional gain of personal participation is less than the private cost.

3.3. The Free Riding Limit on Team Size in an Ongoing Operation

In an ongoing organization, team interaction is repeated over many problem solving cycles. Therefore, the model must include a team member's expectations about the future duration of the ongoing interaction as well as her estimate of how her behavior affects team mates in the future. The expected interaction time is incorporated via a time horizon parameter H determining the discount rate of $e^{-t'/H}$ for a benefit accrued t' time units from the current time t. The shorter the time horizon H, the higher the individual's discount rate for future benefits from further cooperation. We renormalize the time scale t' using a parameter α, which represents the rate at which team members re-examine their cooperation choices. This can be pictured, for example, as the change of behavior resulting from quarterly or semi-annual performance reviews.

Our model starts out with the assumption that employees normally do *not* carry out a game-theoretic analysis of the multi-period interaction within the team. Rather, they are boundedly rational, setting their actions based on simple observations and expectations about the basic nature of colleagues' behavior. We introduce an expectation function $E(f_C)$ to model the individual team member's belief how her behavior influences future behavior of colleagues. Recall that f_C denotes the fraction of team members cooperating. We assume that $E(f_C) = \gamma f_C$, which implies that the individual believes her influence to be higher when more colleagues already cooperate: defection and cooperation alike spread into the rest of the team at rate $E(f_C)$. This form of expectation is referred to as *bandwagon expectations*[9]. It pictures employees who care more about high performers (cooperators) than defectors and only consider the effect of their own behavior on them.

As an example, consider bandwagon expectations in the context of working late in the office. If everyone on the team usually goes home early, then someone who stays longer will not expect others to follow her example -- on the contrary, that person may be seen as an overly ambitious outsider. Once, however, a working atmosphere arises where people often work late, then peer pressure often leads individuals to go along. Moreover, defection by one team member is widely visible and will either be punished by the others or have a significant demotivating effect, leading to spreading defection by colleagues.

[9] See Huberman and Glance [15] for a discussion of alternative forms of expectations.

We now incorporate both time horizon and expectation function into the individual's utility function. With t being the present time, let $\Delta f_c(t+t')$ denote the expected future difference (at time $t+t'$) between the fraction of team members cooperating if the individual cooperates now and if the individual defects now. The expectation of an individual is characterized by three features: First, the individual expects her decision to spread exponentially through the population of colleagues, i.e., collaboration increases as a saturating exponential if she cooperates, and decreases if she defects. Second, the initial slope of this exponential steepens with the current level of cooperation. This is the bandwagon effect described above: the individual believes to have more influence with cooperators than with defectors. Third, the individual believes to be marginalized by large team sizes, i.e., the larger the team, the slower is the saturation of the exponential spreading of the individual's behavior.

Figure 1 shows these characteristics of the expectation function. The upper curve represents how the individual expects cooperation to develop if she cooperates, the lower is the individual's extrapolation of group behavior if she defects. The rate at which the two curves respectively rise and fall is an exponential driven by time horizon and reevaluation rate, with beginning slope determined by the expectation function at the present level of cooperation. The marginalization is incorporated by the exponent having n, the team size, in the denominator: the larger n, the more slowly does the overall cooperation go towards 1 in case of cooperation, or to zero in case of defection.

Figure 1 to be included about here

Summarizing, the form of the extrapolated future cooperation levels become

$$\left.\begin{aligned} \overline{f}_c(t+t') &= 1-(1-f_c)e^{-\frac{\alpha\gamma f_c(t)}{n}t'} \\ \underline{f}_c(t+t') &= f_c e^{-\frac{\alpha\gamma f_c(t)}{n}t'} \end{aligned}\right\}. \qquad (5)$$

We now have the necessary pieces in place to ask how team size is influenced by performance benefits from collaborative problem solving, bonus, time horizon, and expectations. Section 4 will look at the first benchmark case of linear performance benefits, and Section 5 at the second of quadratic benefits.

4. Optimal Team Size for Linear Performance Increase

If $\varepsilon = 0$, team performance increases linearly with the number of fully contributing members. In this case, Equation (5) simplifies, and only the *difference* between future cooperation levels matters:

$$\Delta f_c(t + t') = 1 - e^{-\frac{\alpha\gamma f_c(t)}{n} t'} . \tag{6}$$

Based on this prediction of how they expect future cooperation to evolve, team members decide whether to cooperate or defect based on the expected discounted utility difference. This is the integral of discounted future benefits associated with the extrapolated difference between cooperation levels, Δf_c:

$$\begin{aligned} \Delta U_i(t) &= \int_0^{\infty} [\Delta f_c(t + t')\beta b - c] e^{-t'/H} dt' \\ &= (\beta b - c)H - \frac{\beta b}{\frac{\alpha\gamma f_c}{n} + \frac{1}{H}} . \end{aligned} \tag{7}$$

Team member *i* defects at time *t* if $\Delta U_i(t) < 0$, and cooperates if $\Delta U_i(t) >= 0$. The decision is based on the fraction of the team $f_c(t)$ cooperating at time *t*. Thus, the cooperation decision of the individual team member reduces to the question whether the current level of cooperation is larger than a critical fraction:

$$f_c(t) \geq \frac{nc}{(\beta b - c)\alpha\gamma H} \equiv f_{crit} . \tag{8}$$

This is the first important result of our model: The individual's cooperation decision collapses to a "barrier rule:" *Work hard when overall performance corresponds to a "barrier percentage" of team members working hard, and otherwise shirk.* Individuals are able to simplify a complicated problem of maximizing own benefits to a simple rule requiring only the tracking of overall team performance (and the derived level of cooperation).

In the present context, we are interested in the question for which team size cooperation is sustainable. We can answer this question analytically in our error-free scenario, when employees make no mistakes in trying to contribute and don't misunderstand one another's actions: if $f_{crit} > 1$, it is impossible that cooperation is high enough to sustain itself, and as a result the whole team will shirk. From Equation (7), we get thus a maximum team size for cooperation:

$$n_{max} = \alpha\gamma H\left(\frac{\beta b}{c} - 1\right). \tag{9}$$

If team size is increased to or beyond n_{max}, team members will shirk, and performance will go down to the minimum level (here normalized to zero). Which team size should the manager choose? As long as the team size is smaller than n_{max}, all team members will cooperate. The firm keeps the value of the team's output, minus salaries and wages. Therefore, the firm's profit function resulting from the activity of the team is

$$\Pi_F = n[(1-\beta)b - w] \qquad \text{for} \quad n \le n_{max}. \tag{10}$$

Excluding the trivial case of the firm not gaining enough from the team's activity to pay salaries and bonuses, this profit is increasing in n. Therefore, the manager responsible for staffing the team should choose a team size of n_{max}.

Equation (9) points to several levers through which management can push the cooperative limit upwards. First, of course, the value of team activity b increases the firm's profit as well as team members' motivation (through the bonus). The bonus percentage β boosts team motivation, but at a cost for the firm, as Equation (10) shows. The total effect of bonuses on profit, $\partial\Pi / \partial\beta$, is ambiguous. The individual cost of cooperation c decreases team motivation, justifying work support tools as well as measures building team spirit and a positive atmosphere. A long decision horizon αH also benefits cooperation and firm profit, expressing the value of stability and trust. Finally, cooperation increases with the perceived "influence", γ, that individuals have on the behavior of their colleagues. This means that a tighter team spirit and increased interactions between individuals will reenforce cooperation.

In reality, the increase in solution speed cannot go on forever, for two reasons. First, there is usually a minimum possible solution time. This can be given externally (e.g., by response times from partner groups or companies), or by an internal speed limit: when the number of team members grows to the same order as the complexity of the problem to be solved, T / S (the size of the solution space over the number of possible solutions), then with high likelihood one of them will, by chance, hit a solution right away. The time for going through the heuristic solution procedure once, i.e., finding a solution after the first iteration, corresponds to the maximum possible speed, regardless of further increases in team size. Note that this saturation will only happen for very large teams, since we assume that the problem in question is very complex.

Second, we have ignored the cost of collaboration. Team members must spend time communicating[10], which reduces the time available for problem solving. When team size and the number of hints become very large, either the hints are no longer read or individual productivity suffers. This leads again to a saturation of the team performance.

When we introduce such a performance limit by adding a problem complexity parameter *s*, then the team performance curve from Equation (1) becomes flat at payoff *sb* when it hits the limit *s*:

$$\Pi = \begin{cases} bn_c & n_c \leq s \\ sb & n_c > s. \end{cases} \tag{11}$$

Team members are aware of the performance plateau and take it into account when deciding about their effort level. *When performance (solution speed) has reached bn, no team member will cooperate further*.

To see this, consider a team member reconsidering cooperation at a time t when observed cooperation $f_c(t)$ is so high that $\Pi = sb$, i.e., the performance plateau is reached. In that case, the individual's own cooperation will not boost performance higher (even if others should also decide to increase effort). However, from a defection the individual will still expect the demotivating effect on others. Thus, in evaluating her decision, the individual sees a utility of $(\beta sb / n - c)$ from cooperation versus an exponential decay in team performance, analogously to Equation (6), from defection. Thus, the expected utility difference becomes[11]

$$\begin{aligned} \Delta U_i(t) &= \int_0^\infty \left[\frac{\beta sb}{n} - \beta b f_c e^{-\frac{\alpha\gamma f_c}{n}t'} - c\right] e^{-t'/H} dt' \\ &= \left(\frac{\beta sb}{n} - c\right)H - \frac{\beta b f_c}{\frac{\alpha\gamma f_c}{n} + \frac{1}{H}}. \end{aligned} \tag{11}$$

Note that $\Pi = sb$ implies $nf_c = s$. We can rewrite the utility difference as

$$\Delta U_i(t)\left(\frac{\alpha\gamma f_c}{n} + \frac{1}{H}\right) = \left(\frac{\beta sb}{n} - c\right)\frac{H\alpha\gamma f_c}{n} - c < 0, \tag{12}$$

since we are assuming existence of a dilemma. This implies immediately that *no team member will choose cooperation* when performance saturation is reached ($f_c(t)$ is such that $\Pi = sb$).

---

[10] In the context of the simulation experiments, members need time to write and read hints on the blackboard.
[11] For simplicity of exposition, we drop the t in $f_c(t)$ whereever the context is clear.

We conclude that the manager responsible for the organizational unit should set team size to the limit where team motivation breaks down due to either free riding or saturation of performance from collaboration. The additional parameter s is an expression of the problem complexity faced by the team: the more complex the problem, the more workers can meaningfully collaborate on it. Our analysis implies that problem complexity, even if it is valuable, should be limited to the extent that the associated solution team remains small enough to support team atmosphere and high motivation.

We now turn to the situation of high benefits from collaborative problem solving. If team performance increases quadratically with team size, then it balances motivational effects: the payoff of working hard is so high that the social dilemma does not worsen with a largening group. Only the externally given performance saturation limits team size.

5. Optimal Team Size for Quadratic Performance Increase

If $\varepsilon = 1$, i.e., team performance increases *quadratically* with the number of fully contributing members, we can evaluate the individual's expected discounted utility difference between cooperation and defection analogously to Equation (7):

$$\begin{aligned}\Delta U_i(t) &= \int_0^\infty \{\beta bn[(1-(1-f_c)e^{-\frac{\alpha\gamma f_c(t)}{n}t'})^2 - f_c^2(e^{-\frac{\alpha\gamma f_c(t)}{n}t'})^2] - c\}e^{-t'/H}dt' \\ &= \beta bn[H - \frac{2(1-f_c)}{\frac{\alpha\gamma f_c}{n}+\frac{1}{H}} + \frac{1-2f_c}{\frac{2\alpha\gamma f_c}{n}+\frac{1}{H}}] - cH\end{aligned} \tag{13}$$

We can again consider the limiting situation where the individual is indifferent between cooperation and defection, and solve the resulting quadratic equation for a critical value of f_c. This quadratic equation has only one positive root, which implies that the individual will cooperate as long as:

$$f_c(t) \geq \frac{\frac{3}{4}c + \sqrt{(\frac{3}{4}c)^2 + \frac{1}{2}c\beta bn[(1-\frac{c}{\beta bn}) + \frac{n}{\alpha\gamma H}]}}{\beta b\alpha\gamma H[(1-\frac{c}{\beta bn}) + \frac{n}{\alpha\gamma H}]} \equiv f_{crit}. \tag{14}$$

We find that the individual still follows a barrier policy, as in the linear case. The individual is still able to simplify the complicated decision problem, setting her effort based only on the inferred cooperation level.

We ask again for which team size cooperation is sustainable, or for which n we have $f_{crit} < 1$. Inspection of Equation (14) shows that f_{crit} is not monotonely increasing in team size. Its change over n follows the pattern shown in Figure 2, and it approaches a well-defined limit:

$$\lim_{n \to \infty} f_{crit} = \sqrt{\frac{c}{2\beta b \alpha \gamma H}}. \tag{15}$$

Figure 2 shows the shape of f_{crit} as a continuous function of team size (ignoring the fact that in the model only integer values of n are relevant). Only the interesting case is shown, where the limit of the function is below 1. The limit above 1 indicates that team economics are not viable and that cooperation is unsustainable. The shape of the curve depends on relative parameter sizes. Numerical experiments have demonstrated that the function monotonely decreases if the peer influence parameter γ is low and that the function "overshoots" if γ is high. The initial decrease of f_{crit} stems from low team performance associated with a small number of cooperating team members. Performance rapidly increases with more workers cooperating, allowing for a lower fraction of cooperation.

Figure 2 to be included about here

The difference between f_{crit} for high and low γ (the "overshoot" of the high influence curve) is caused by the difference in speed of an individual defection's spread through the population (see Figure 1). If cooperation is already high, then fast spreading defection causes a much higher discounted loss to the individual than a slow spread, whereas the upside is similar for both speeds, since cooperation cannot go much higher. Therefore, cooperation is more advantageous for a high influence parameter γ (fast spread of the individual's action), resulting in a lower f_{crit}.

We conclude from this analysis that quadratic collaboration benefits are able to balance the negative motivational effects of team size (i.e., the increasing temptation of shirking). The cooperation barrier f_{crit} does not necessarily approach one with increasing team size, and team cooperation is sustainable as long as the quadratic performance increase can be maintained.

As in the linear case, there is in reality a performance maximum inherent in the problem solved by the team, and coordination costs will also prevent further increases when the team gets very large. The point here is that, in contrast to the linear case, team size is limited *not* by motivational effects, but only by structural limits on the ability to maintain the quadratic growth of solution speed. As in Section 4, it can be shown that an upper performance limit cuts cooperation by the team members, and that it is advantageous for the firm to actually increase team size up to that point.

This result points to highly productive problem solving processes as a way of offering incentives to the employees to overcome their personal cost of effort. *Effective problem solving methods are of double importance : They do not only directly increase team productivity, but can also help to mitigate the social dilemma of team production.*

6. Summary and Outlook

In this paper we have presented an analytical model of organizational size. In contrast to existing models of firm size in the Industrial Organization literature, our model is not formulated in terms of high level constructs (such as "loss of control"), but trades off two phenomena that influence day-to-day operations of an organization. The two phenomena are problem solving power gained through collaboration in teams, and free riding when groups get large. Thus, our model has concrete managerial interpretations and can be applied to several layers in a firm, from the working team to larger structures.

In particular, we introduce a concrete mechanism of team collaboration, through parallel activity combined with frequent sharing of partial information about a problem solution. This mechanism of collaborative problem solving leads to a performance increase, measured as solution speed, that is between linear and quadratic as a function of team size. The formulation includes explicit representations of problem complexity, quality of problem solving methods, and effectiveness of collaboration. The model is thus applicable to many situations faced by teams in practice, from engineering design to strategy formulation.

The team is paid a linear performance bonus shared equally by all team members. Each team member periodically sets her effort level either to contribute at the best of her ability (to "cooperate") or to "cruise" at the minimum level not recognizable as shirking (to "defect"). The team members' individual effort level is unobservable by colleagues or managers. An individual's decision about cooperation is boundedly rational, based on rewards (bonuses), the time horizon of team interaction and thus the size of future benefits, and by the individual's expectations about colleagues' behavior. We model expectations explicitly as "bandwagon expectations": first, the individual believes that her own actions matter less in a larger group, and second, the influence of a single person's attitude on her peers increases with the general level of team cooperation.

In this situation, the individual's cooperation decision collapses to a simple "barrier rule:" Work hard when a certain "barrier percentage" of team members work hard, and otherwise shirk. The influence of team size on the barrier percentage depends on the benefits of collaboration (i.e., on problem solving methods and collaboration effectiveness): When performance grows linearly with cooperating team size, then the barrier percentage increases with team size and eventually grows beyond 100%. Thus, cooperation is only sustainable up to a certain team size. When benefits grow quadratic in the number of cooperating team members, then they balance the temptation to shirk. In this case, team size is only limited by an exogenous performance saturation, such as communication costs or a performance maximum. This result points towards the double value of effective problem solving and collaboration: it increases direct productivity *and* mitigates the social dilemma inherent in team production.

Implicit in this model formulation is the assumption that more complex problems are more valuable for the team to work on, or in other words, value created is proportional to time spent. That only makes sense as long as it is assured that problems are defined "efficiently", i.e., that they are not made more complex than necessary. This assumption is restrictive and a consequence of our choice in this paper to present the model in a simple version with close form analytical exposition. Problem complexity independent of solution value, communication cost, and as other operational decision parameters can be explicitly incorporated in richer extensions of the model.

While our model is sufficiently simple to allow analytical solutions, it includes a number of operational parameters of managerial interest: problem complexity, solution value, bonuses, team horizon and stability, individuals' cooperation cost, and expected influence of behavior on others. Thus, the model sheds, even in the present simple form, some prescriptive insight on the drivers of organizational unit size.

In addition, the basic theory outlined in this paper can be expanded to include many dimensions of organizational design: Performance measures can be cost or output quality in addition to solution time. In modeling collaborative problem solving, it is possible to include the quality of solution procedures, quality of information exchange, organizational learning and memory, and organizational structure. One can even model the influence of process structure on problem solving, such as the possible extent of parallelism and the associated communication effort, or the necessity to iterate in a multi stage project because of the quality of partial solution handed over from one stage to the next. This is highly relevant to modeling concurrent engineering, to name one example. In modeling team motivation, it is possible to include employee training, company work ethic, interactions of several organizational layers and units, and organizational changes (transfers and reorganizations).

Simple closed form analytical solutions as in the present paper will no longer be possible in richer and generalized versions. However, dynamic simulations of system behavior are possible and have already been built for certain aspects (e.g., [12]). Simulations can build an understanding of behavioral characteristics in organizations, and they can represent the dynamic evolution over time in addition to the steady state "average" behavior described in the present paper.

Barron [4] calls for further development of theoretical models of organizations, on which empirical research can be based. Our model offers explicit representations of collaborative problem solving and motivation, and thus represents a step towards a *Theory of Organizational Processes*. Such a theory should contribute to knowledge on the performance of operational processes *combined with* people motivation and interaction. This reflects our belief that management science ultimately needs to bridge the separate bodies of knowledge in organizational behavior on one side and engineering and operations management on the other.

References

[1] A. A. Alchian and H. Demsetz, "Production, Information Costs, and Economic Organization," *American Economic Review* **62**, 1972, pp. 777 - 795.

[2] Th. Allen, *Managing the Flow of Technology,* MIT Press, 1977.

[3] H. L. Angle, "Psychology and Organizational Innovation," in *Research on the Management of Innovation: The Minnesota Studies,* A. H. Van de Ven, H. L. Angle, and M. S. Poole, eds., Harper & Row, 1989, pp. 135 - 170.

[4] T. Barron, "Impacts of Information Technology on Organizational Size and Shape: Control and Flexibility Effects," *Journal of Organizational Computing* **3**, 1993, pp. 363 - 387.

[5] J. Bendor and D. Mukherjee, "Institutional Structure and The Logic of Ongoing Collective Action," *American Political Science Review* **81**, 1987, pp. 129 - 154.

[6] D. S. Bree, "The Distribution of Problem Solving Times: An Examination of the Stages Model," *Br. J. Math. Stat. Psychol.* **28**, 1975, pp. 177 - 200.

[7] E. Brynjolffson and H. Mendelson, "Information Systems and the Organization of the Modern Enterprise," *Journal of Organizational Computing* **3**, 1993, pp. 245 - 255.

[8] E. Brynjolffson, T. W. Malone, V. Gurbaxani, and A. Kambil, "Does Information Technology Lead to Smaller Firms?" Working Paper, Center for Coordination Science, MIT, 1989.

[9] Calvo, G. A., and S. Wellisz, "Supervision, Loss of Control, and the Optimum Size of the Firm," *Journal of Political Economy* **86**, 1978, pp. 943 - 952.

[10] S. H. Clearwater, T. Hogg, and B. A. Huberman, "Cooperative Problem Solving," in *Computation: The Micro and the Macro View,* Huberman, B. A., ed., World Scientific, 1992, pp. 33 - 70.

[11] N. S. Glance and B. A. Huberman, "The Outbreak of Cooperation," *Journal of Mathematical Sociology* **17**, 1993, pp. 281 - 302.

[12] N. S. Glance and B. A. Huberman, "Organizational Fluidity and Sustainable Cooperation," in *Computational Organization Theory,* M. Prietula and K. Carley, eds., Erlsbaum Publishing, 1994, pp. 217 - 240.

[13] N. S. Glance and B. A. Huberman, "Social Dilemmas," *Scientific American* , March 1994, pp. 58 - 63.

[14] B. A. Huberman, "The Performance of Cooperative Processes," *Physica* **42**, 1990, pp. 38 - 47.

[15] B. A. Huberman and N. S. Glance, "Beliefs and Cooperation," in *Modelling Rational and Moral Agents,* P. Danielson ed., Oxford University Press, 1995.

[16] M. Keren and D. Levhari, "The Internal Organization of the Firm and the Shape of Average Costs," *The Bell Journal of Economics* **14**, 1983, pp. 474 - 486.

[17] S. Martin, *Advanced Industrial Economics,* Blackwell, Oxford, 1993.

[18] P. Milgrom and J. Roberts, "Economic Theories of the Firm: Past, Present, and Future," *Canadian Journal of Economics* **21**, 1988, pp. 444 - 458.

[19] J. A. Mirrlees, "The Optimal Structure of Incentives and Authority Within an Organization," *The Bell Journal of Economics* **7**, 1976, pp. 105 - 131.

[20] O. E. Williamson, "Hierarchical Control and Optimal Firm Size," *Journal of Political Economy* **75**, 1967, pp. 123 - 138.

[21] O. E. Williamson, *The Economic Institutions of Capitalism,* New York, The Free Press, 1985.

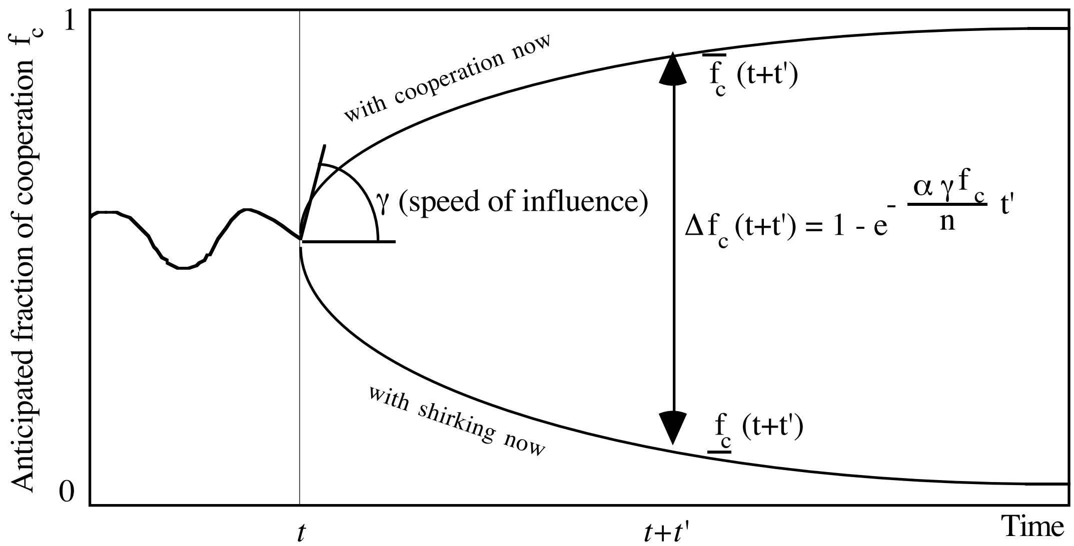

Figure 1: Bandwagon Expectations About Future Cooperation

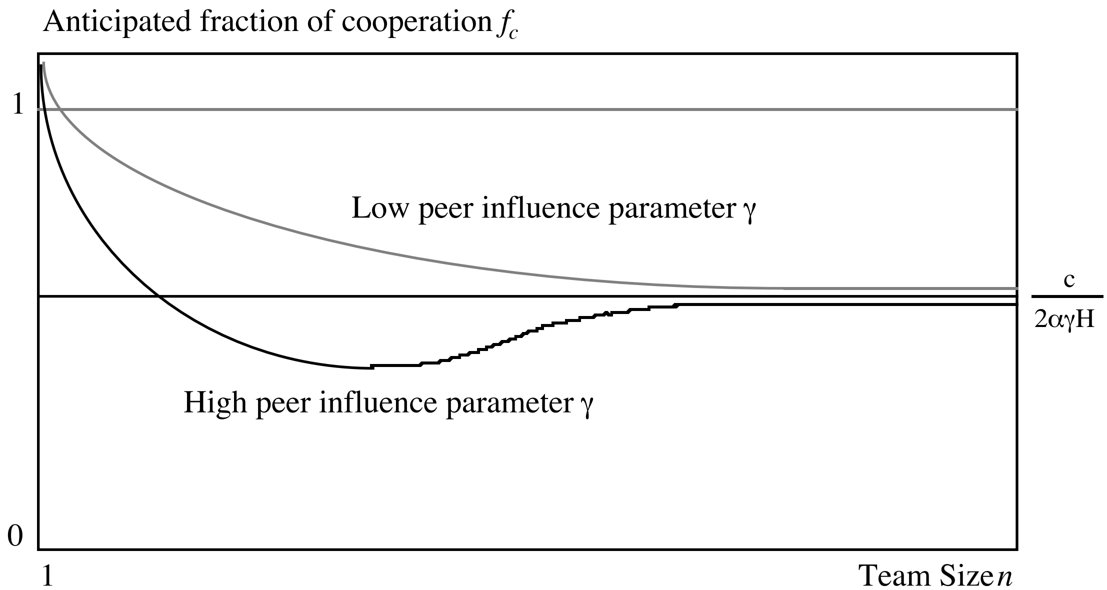

Figure 2: The Cooperation Barrier as a Function of Team Size