What individuals don't know the collective does
A possible encoding for collective information

Fernando Esponda, Department of Computer Science, Instituto Tecnológico Autónomo de México

1. INTRODUCTION

In this paper I address the issue of collective intelligence from the standpoint of how information is sometimes represented and stored among the participating agents. In particular I propose that for a given trait or property the information describing it might not be accesible in its entirety to all the participating entities, that it might be distributed amongst them and codified negatively. I discuss two examples from nature in which this mechanism seems to be at work and give two applications from computer science and statistics that exploit it.

2. T CELLS

The first example comes from the vertebrate immune system. One of the tasks the immune system (IS) is responsible for is maintaining the body free from pathogens—foreign agents that can cause disease. The problem the IS faces is that of how to recognize novel pathogens or pathogens that have evolved sufficiently as to not resemble some previously encountered predecessor. Part of the answer comes from the T cells maturation process: T cells—a special kind of immune cell—are an important component for mounting an immune response, their role is to patrol the body and monitor the surface of cells for evidence of pathogens. They originate in the bone marrow with an untuned capacity to detect proteins in the surface of cells and migrate to the thymus where they undergo a process of selection. The thymus is an organ that contains cells that express a sample of the proteins normally found elsewhere in the body and is presumably free from disease. T cells patrol it, as they would the body, but die here if they mistakenly identify an antigen, a process known as negative selection [Sompayrac 1999]. Those T cells that survive are released into the body and will likely not recognize *self* proteins, whatever they do recognize is *non-self* and a potential pathogen. Each T cell is thus capable of identifying a subset of *non-self* and together all T cells can identify most of it. We could say that the collection of all such cells define what *self* is by individually specifying what it is *not*. This strategy keeps T cells and their creation simple and allows for the immune system to track a movable definition of *self* by translating its changes into changes of only some T cells.

3. ANTS

Following this idea recent work [Esponda and Gordon 2015] proposes that a similar mechanism might be at work in ants and other social insects. Ants are a kind of super organism in which individuals of a colony work in concert to maintain the colony itself. One fundamental capacity ants must have to preserve nest cohesion is the ability to tell nest mates from non-nestmates. Whenever two ants meet they sense each other's cuticles, which are covered in hydrocarbons, in search for cues that can help them make that determination—much like T cells search the surface of cells for telling signals. Also, analogous to T cells in the immune system, ants are faced with the task of recognizing odors (cuticular hydrocarbon combinations) which they probably haven't encountered before; for this, nestmate recognition models [Sturgis and Gordon 2012] assume that ants take as standard known colony properties. The model described in [Esponda and Gordon 2015] suggests that ants become habituated to their nestmate's odors and that each individual is capable of identifying an incomplete and unique set of non-nestmate odors; it further posits that this classification ability can adapt to reflect the particular life histories of each individual. The characteristic all ants share is that none of them are likely to react to their nest's odors: non react to *self*, and they differ in which elements of *non-self* they are able to identify. In this way the collection of ants of a nest define its identity by individually recognizing some of what it is *not*. The benefit of having identity distributed in this way is the possibility to dedicate less resources per ant for this endeavor and the agility to track the change in odors that a colony constantly undergoes[Vander Meer et al. 1989].

4. SECURITY

A natural question to ask is if the idea of having information encoded negatively has any utility for digital data. One possibility, described in [Esponda et al. 2009], follows the notion that both ants and T cells perform a security service to their collective and do so using the distributed, negative representations described above. It suggests that perhaps by encoding data negatively, say from a file full of secrets, it can also acquire some security properties. The proposal is that given a database (more precisely a list of entries) a negative version of it can be created in which each record describes some of what is *not* in the original and collectively delineate its contents. The first question is how to create this negative database without incurring an exponential need for storage, clearly the number of possible entries for a typical list, with fixed length entries, is much much smaller than what is actually there: imagine the possible names and addresses that are not in your address book. The strategy is to have each negative database entry specify a large subset of what the true database does *not* contain, in analogy to the large subset of non-nestmate odor that an ant can recognize and to the antigens that a T-cell can bind to. In essence the negative database creation algorithm is much like a compression program except that instead of outputting *the_file.zip* it outputs *everything_but_the_file.zip*. A second question is how this can be used for security. One answer is that it will increase the security of the original database if each negative record is kept in separate places, since a single negative record contains limited information. Another answer has to do with how the "compression" algorithm works. If done right, the process of inferring a single positive record from the negative database is an $\mathcal{NP}$-Complete problem, meaning there is no known efficient algorithm to accomplish this—interestingly, this also means that asking whether a particular positive record is in the database can be answered promptly! This result points towards a possible connection between the theory of computation and the recognition capabilities of both the immune system and ant colonies.

5. PRIVACY

Finally, the concept of there being things known by the collective that are ignored individually, the wisdom of the crowds [Galton 1907; Surowiecki 2005], can be leveraged for situations in which there is a collective property of some interest, say the income distribution of a group of people or the average speed on a highway, but where we wish the individual traits to remain obscure. The work presented in [Esponda and Guerrero 2009; Esponda et al. 2015] proposes an instrument for gathering information that allows each respondent to retain some uncertainty regarding the polled trait while still collecting enough information to compute meaningful population statistics. This instrument, called a Negative Survey, turns the idea of an agent being able to identify only part of *non-self* to that of a responding agent revealing only some information (negatively encoded) regarding its true answer.

Consider a standard questionnaire in which there is a single question and n exhaustive and mutually exclusive possible answers. A queried respondent would then choose one and only one option and reveal his answer. A negative version of this questionnaire would have the question negated and $n-1$ possible answers (see Fig. 1). The queried respondent can now choose one of the $n-1$ options

The total amount of your assets is:

a) Between 0 and 20,000
b) Between 20,001 and 50,000
c) Between 50,001 and 70,000
d) 70,001 or more

(a) Positive question

The total amount of your assets is **not**:

a) Between 0 and 20,000
b) Between 20,001 and 50,000
c) Between 50,001 and 70,000
d) 70,001 or more

(b) Negative question

Fig. 1. A positive question and its negative version.

and, provided n is greater than 2, be certain that he has revealed less information than in the positive counterpart. Interestingly, if the survey includes a big enough sample it is possible to accurately compute the frequency distributions, not only for the negative survey, but of its positive counterpart as well. The missing information withheld by each individual is compensated by knowing how each respondent chooses among the available $n - 1$ options. In the simplest case respondents are asked to choose their answer uniformly at random so that each of the available options is equally likely to be selected. For example, consider the negative question of Fig. 1(b) and suppose 100 people are queried, that 20 answer *a*, 30 answer *b*, 20 answer *c*, and that 30 answer *d*. Let P_a, P_b, P_c and P_d be the number of respondents positively belonging to each category, which we ignore. Given the above data we can estimate P_a, for instance, by noting that the total number of people that chose *a* negatively is composed of some of the people that positively belong to *b*, some of the people that positively belong to *c*, and some of the people that positively belong to *d* : $20 = \frac{1}{3}P_b + \frac{1}{3}P_c + \frac{1}{3}P_d$, where the $\frac{1}{3}$ is the probability that someone whose positive category is *b*, *c* or *d* choses *a* as its negative answer. Thus $20 = \frac{1}{3}(P_b + P_c + P_d) = \frac{1}{3}(100 - P_a)$ and solving for *a* we can estimate of P_a as $\hat{P_a} = 3\frac{100}{3} - 20 = 40$. This example illustrates that at least for some issues, asking direct questions is asking for more information than what is actually needed and that population statistics can be computed without the necessity of imputing specific answers to specific individuals.

ACKNOWLEDGMENTS

I gratefully acknowledge the support provided by Asociación Mexicana de Cultura, A. C.

REFERENCES

Fernando Esponda, Stephanie Forrest, and Paul Helman. 2009. Negative representations of information. *International Journal of Information Security* 8, 5 (2009), 331–345.

Fernando Esponda and Deborah Gordon. 2015. Distributed Nesmate Recognition in Ants. *In revision* (2015).

Fernando Esponda and Victor M Guerrero. 2009. Surveys with negative questions for sensitive items. *Statistics & Probability Letters* 79, 24 (2009), 2456–2461.

Fernando Esponda, Kael Huerta, and Victor Guerrero. 2015. Collecting private information: A suitable proposal for Big (sensitive) Data. *In revision* (2015).

Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature* 75 (1907), 450–451.

Lauren Sompayrac. 1999. *How the Immune System Works*. Blackwell Science.

Shelby J Sturgis and Deborah M Gordon. 2012. Nestmate recognition in ants (Hymenoptera: Formicidae): a review. *Myrmecol News* 16 (2012), 101–110.

James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

Robert K Vander Meer, David Saliwanchik, and Barry Lavine. 1989. Temporal changes in colony cuticular hydrocarbon patterns of *Solenopsis invicta*. *Journal of Chemical Ecology* 15, 7 (1989), 2115–2125.