Arguments against the Troll (position paper)

Paolo Torroni and **Marco Prandini** and **Marco Ramilli**
University of Bologna, Italy

João Leite
CENTRIA, Universidade Nova de Lisboa, Portugal

João Martins
CENTRIA, Universidade Nova de Lisboa, Portugal
& Carnegie Mellon University, USA

Abstract

We envision an improved social Web, in which the Trolls' disruptive power is inhibited or restricted, and the content produced by and shared among community members can gain authoritativeness. We believe that argumentation theories have the potential to give a key contribution to this vision. We sketch a research path in this direction and discuss some research questions.

It can be argued that the Web 2.0's key to success resides in participation. 500 million active FaceBook users exchange 30 billion content items every month.[1] Wikipedia pages and registered users are counted by tens of millions, page edits by hundreds of millions.[2] No doubt social computing is a hype. Probably it is over-hyped. But it is clearly changing the way people interact, business is run, services are provided, information is produced and shared, Internauts spend their virtual life, real people make friends and real ties are broken up. We could speak of a large scale cultural, social and technological revolution.

Of the innumerous questions that come together with such a revolution, one is particularly interesting to us. That has to do with authoritativeness of social Web content, and ultimately with the survival of the social Web phenomenon itself as we know it today. The Web 2.0 may be the expression of freedom of speech and democratic participation. This is all very good and welcome, but: is it going to last? What will be the face of tomorrow's social Web? Today's Web 2.0 is participated by different groups of users with very different backgrounds and attitudes. We find many of those eager to share their family secrets, joys and pains with the rest of the world; the IT-savvy, wary about spilling their life to the social Web and about relying on information produced by it; those planning their holidays based on TripAdvisor ratings; the lovers of solitude, who finally found a peaceful beach and are going to give it very low marks on TripAdvisor, in the hope to lie there alone for a little while longer. These are the crowds that will elect the next president of the United States.

[1] Source: `http://www.facebook.com/press/info.php?statistics`.

[2] Source: `http://en.wikipedia.org/wiki/Special:Statistics`.

Nice. And problems come along. The Web 2.0 has given rise to a good number of emerging phenomena such as social stalking, multiplication of junk sources and time sucks and gossip amplification. Privacy is probably at its historical minimum. One common reason, if not for all at least for many of these negative effects of the social Web, is the lack of authoritativeness in the Web content. In the social Web, everything could be equal to everything else. The very idea of grassroots and diffused democracy, that underlies the Web and its social version of today, is the reason of this lack of authoritativeness of the social Web contents. We can't do anything about it!

Or, can we?

What is a Troll?

According to a most authoritative source,[3]

> A *Troll* is someone who posts inflammatory, extraneous, or off-topic messages in an online community, such as an online discussion forum, chat room, or blog, with the primary intent of provoking other users into a desired emotional response or of otherwise disrupting normal on-topic discussion.

Online discussion forums are a paradigm of how the social Web functions. Consider for instance some of the best-known gadgets in the market: FaceBook's Comment and Like/Unlike features. Interaction among social community members nowadays takes the form of a series of comments posted below one another, and popularity, which users care so much about, is a function of the number of connections, comments, and likes. A sort of folkloristic H-index. E-mail, which being a private thing is outside the Web 2.0, is less often used now. People find it more efficient and rewarding to use their time by twitting their thoughts to their world of followers, instead of pushing personal messages into mailboxes of users who have less and less time to enjoy them, and who may not even read them or care after all.

Social networking platforms have thus become a sort of factories of mass opinion. Users post their comments, and the crowd in many cases converges to some better focussed mainstream positions with a lot of likes. "Mourinho is a

[3] Source: `http://en.wikipedia.org/wiki/Troll_(Internet)`

great coach!" "He is arrogant." "He is smart." "Inter does not deserve him." "Real Madrid rules." "C'è poco da stare Allegri."[4]

Now, what happens if some comments are posted exactly with the intent of disrupting normal on-topic discussion? This is actually what happens in many politically-oriented forums and blogs, where we find all sorts of posts, most of them on-topic, but some of them intentionally aggressive, racist, uninformative, utterly contradictory or simply irrelevant. More often than not, as a consequence of such comments, the focus of the discussion shifts, at least temporarily, to easy criticisms to these posts, and the main topic of discussions and thread of thoughts collected so far is lost. If not, typically the troll which has been ignored insists until he reaches the objective of depriving an interesting discussion of its initial meaning and results. Our Troll is not simply posting an *ad hominem*: he is effectively running an attack on the authoritativeness of the whole undergoing discussion.

The problem is so relevant that a lot of research in social communities is being carried out to understand how to harness this phenomenon. For example, Kiesler et al. focus on how to specify guidelines and rules, in the belief that when norms are clearly stated, people are more likely to act consistently with the norm, and do so over a wide variety of situations (Kiesler et al. In press).

Instead of specifying norms and rules (which could be done anyway), we propose to augment social networking platforms with means to keep discussions on-topic and let strong positions emerge in a clear way, robustly holding a prominent location in the forum. In other words, we think that keeping Trolls at bay and fighting against their invasion of sensible discussion groups should be the responsibility of the platform, not of the participants in the discussion.

Let us see how this can be done.

Towards an authoritative social Web

We can identify two schools of thought about coping with the lack of authoritativeness of social Web content. One is just accepting the problem. The Web is a source of falsehood, there is no guarantee on the accuracy of data and sources, etc., so Web content should be licensed "as is" - so to say. Another one is reproducing on the Web the classic reputation-based mechanism of traditional press publishing. An article is authoritative, some Web content is true and reliable, because it was published on some renown Web site with a good reputation.

Both approaches have apparent advantages and shortcomings. The first one promotes openness, but it does not address the problem; the second one does not promote openness, but it provides a partial solution to the problem in a controlled environment. Probably these ways of doing content sharing are both necessary for a healthy life of the Web. But is there a third way, which goes in the direction of combining openness with trustworthiness of Web content?

Recent advances in automated text extraction and, specifically on Web dispute identification (Ennals, Trushkowsky, and Agosta 2010), lead us to believe that soon the technology will be ripe to identify claims in discussion forums effectively and automatically, or at least semi-automatically (e.g. with the help of the social community). With this assumption in mind, the problem of keeping an ongoing discussion on-focus is reduced to the problem of isolating claims and presenting them in a way that maximally emphasizes on-topic claims, highlighting the relations between them, and maximally hides off-topic claims.

In the absence of a moderator, and following the democratic spirit of the Web 2.0, what is on and what is off should be defined dynamically, based on inputs that should come from the participating crowd. For example, we could define some *degrees of authoritativeness*, to be assigned by experts, social Web followers, or even by the occasional browser. And of course we should accept that the Troll will also delight himself by assigning his own scores to contents. The mechanism shoud be designed in a robust enough way to cope with a limited amount of Trolls.[5]

If discussions were presented in such a way, the authoritativeness of an online discussion could be measured as a function of its being on-focus and by the support it gained from the crowd (e.g., number/fraction of users agreeing with some claim). Discussions would still be open.

Could this work? At least, we would like to try and see. The success of such a solution would have to be measured empirically.

Argumentation in the social Web

Let us sketch the possible functioning of what we will call "**A***uthoritativ***e So***cial Web* **P***latform*" (**AeSoP**).

In order to be successful, **AeSoP** will first of all need to be accepted by the social Web users. We can isolate some key factors that made the Web 2.0 become so popular.

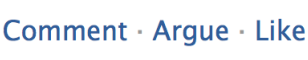
- Large-scale availability of the technology. This is also due to social networking platforms' openness to third-party applications.
- Simplicity. Successful social computing services are based on few mechanisms, which are already known to the user, or easy to be learned.
- Low cost. The user does not have to pay, or pays very little, for social Web services.
- Entertainment. In many applications, information and social exchange has a ludic connotation.

We think that it would be a good idea to stick to these guidelines when designing **AeSoP**. A first release could be thought for a concrete existing social networking platform, such as FaceBook (FB). Following the simplicity cri-

[4] Only for soccer followers: the sentence literally means *Not much left to be Cheerful for*. "Cheerful" (Allegri) is the name of AC Milan's coach. This comment was posted when Mourinho's Real Madrid scored the second goal against Allegri's AC Milan in two consecutive minutes in a 2010 Champions League match. The position genially summarized by such a caustic comment received many likes.

[5] When Trolls gain the majority, in Democracy they become the good guys and they are expected to take charge of the system. Such is (democratic) life.

terion, **AeSoP** could be integrated into FB's well-known basic mechanisms, such as Comment, used to post a comment about a given status, comment, link, etc. (social content item), and Like, to express agreement with some item.[6]



Besides Comment and Like, a third possibility could be Argue: a starting point for a focussed discussion about a certain item. The Argue option will allow the user to post comments in favor or against a claim identified inside the item. The user that initially produced the item can decide whether to accept the challenge. More social users can join the discussion.

As the discussion proceeds, **AeSoP** will use colors, fonts, geometries or other visual artifacts to highlight a prevailing opinion, and emphasize agreements, supporting arguments, attacks and contradictions. The logic behind **AeSoP** would be a mechanism to combine an argumentation framework with community feedback. The idea is to extract a set of claims, which we can call "socially acceptable" arguments, to isolate the positions of the community with respect to the matter under discussion. A simple type of community feedback would be votes, such as FaceBook's likes, or bipolar votes such as likes and dislikes. Suitable candidates for the basic argumentation framework could be Dung's abstract argumentation framework (1995), Bench-Capon's value-based argumentation frameworks (2002), Dung, Kowalski & Toni's assumption-based argumentation frameworks (2009), Modgil's meta-level argumentation frameworks (2009), Cayrol & Lagasquie-Schiex's bipolar abstract argumentation systems (2005), Baroni, Giacomin & Guida's strongly connected components-based argumentation semantics (Baroni, Giacomin, and Guida 2005), or a combination of them. Votes could be attached to arguments as proposed by Martins (2010), or they could be recorded in an argumentative way themselves, for instance by making use of argumentation schemes (Walton, Reed, and Macagno 2008).

There are already many research results that we can use for the realization of **AeSoP**. Besides existing argumentation theories and implementations, concepts taken from the above mentioned Dispute Finder project (Ennals, Trushkowsky, and Agosta 2010) could be used to find and highlight disputed statements in discussion group comments. Existing visualization tools (Kirschner, Shum, and Carr 2003) could be used to enhance clarity of presentation and promote user acceptance.

Discussion

The use of argumentation has already been proposed before in relation with the Web 2.0 (Torroni, Gavanelli, and Chesani 2009), and now the topic has gained a lot of attention also from the Semantic Web community (Schneider et al. 2010). Our proposal however poses new research questions.

Such a system should be as self-regulated as possible. The need for third-party interventions would jeopardize scalability and would contradict the purpose of fostering authoritativeness through democratic debate of the issues.

The envisioned regulatory forces are of two kinds: user-provided annotation, and automatic text analysis.

The former mechanism leads us to the first open question. Is it plausible to build a system that relies only on the prevalence of honest contributions to converge to the desired outcome?

The feasibility of the latter mechanism is an open question in itself. Would the system be able to extract and compare the meaning of user contributions by means of text extraction techniques, maybe by learning from previously validated argumentation structures? Such a behavior would have very interesting applications. Each new contribution could be classified accordingly to a taxonomy leading to various possible effects, for example: automatically placing the contribution within the argument, suggesting the most likely placement to users, flagging the contribution as harmless but undecidable, flagging the contribution as potentially harmful, or automatically discarding it. Would this kind of mechanism be applicable to contexts different from Web 2.0 as well, like for example to spam filtering?

Another research question that we want to address is about the convergence of Computational Argumentation, Philosophical Argumentation, Philosophy of Language, Cognitive Sciences, English Composition, Critical Thinking and other relevant branches of philosophy, language studies and social sciences. We believe that **AeSoP** tools could be used by scholars interested in analyzing the social behavior of arguing crowds and the evolution of thinking under the bias of cooperation technologies. We also believe that **AeSoP** could be a very useful tool to compare the many semantics of computational argumentation proposed in the literature, and see how well they reflect the intuitions of human thinking.

We are currently investigating theoretical issues, to define a model of argument and discussions, the relations between arguments and the semantics of social argumentation in **AeSoP**. At the same time, we are investigating technological issues, working on the requirements, and exploring the exiting technologies. All suggestions are welcome.

[6] Actually, the semantics of Like is not at all clear, and people use it in most incoherent ways, to express sometimes agreement, some other times sympathy, joy, sorrow, contempt, amazement, or simply to indicate that a certain item deserves attention.

References

[Baroni, Giacomin, and Guida 2005] Baroni, P.; Giacomin, M.; and Guida, G. 2005. Scc-recursiveness: a general schema for argumentation semantics. *Artif. Intell.* 168(1-2):162–210.

[Bench-Capon 2002] Bench-Capon, T. J. M. 2002. Value-based argumentation frameworks. In Benferhat, S., and Giunchiglia, E., eds., *NMR*, 443–454.

[Cayrol and Lagasquie-Schiex 2005] Cayrol, C., and Lagasquie-Schiex, M.-C. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In Godo,

L., ed., *ECSQARU*, volume 3571 of *Lecture Notes in Computer Science*, 378–389. Springer.

[Dung, Kowalski, and Toni 2009] Dung, P.; Kowalski, R.; and Toni, F. 2009. Assumption-based argumentation. In Rahwan, I., and Simari, G., eds., *Argumentation in AI*. Springer. 199–218.

[Dung 1995] Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77(2):321–358.

[Ennals, Trushkowsky, and Agosta 2010] Ennals, R.; Trushkowsky, B.; and Agosta, J. M. 2010. Highlighting disputed claims on the web. In Rappa, M.; Jones, P.; Freire, J.; and Chakrabarti, S., eds., *WWW*, 341–350. ACM.

[Kiesler et al. In press] Kiesler, S.; Kittur, A.; Kraut, R.; and Resnick, P. (In press). Regulating behavior in online communities. In Kraut, R. E., and Resnick, P., eds., *Evidence-based social design: Mining the social sciences to build online communities*. Cambridge, MA: MIT Press.

[Kirschner, Shum, and Carr 2003] Kirschner, P. A.; Shum, S. J. B.; and Carr, C. S., eds. 2003. *Visualizing argumentation: software tools for collaborative and educational sense-making*. London, UK: Springer-Verlag.

[Martins 2010] Martins, J. 2010. Argumentation systems with social voting. Master's thesis, Mestrado em Engenharia Informática, CENTRIA, Departamento de Informática, FCT, New University of Lisbon, Portugal, Quinta da Torre 2829-516 Caparica, Portugal.

[Modgil 2009] Modgil, S. 2009. Reasoning about preferences in argumentation frameworks. *Artif. Intell.* 173(9-10):901–934.

[Schneider et al. 2010] Schneider, J.; Passant, A.; Groza, T.; and Breslin, J. G. 2010. Argumentation 3.0: how Semantic Web technologies can improve argumentation modeling in Web 2.0 environments. In *Proc. 3rd COMMA, Desenzano del Garda, Italy*. IOS Press.

[Torroni, Gavanelli, and Chesani 2009] Torroni, P.; Gavanelli, M.; and Chesani, F. 2009. Arguing on the semantic grid. In Rahwan, I., and Simari, G., eds., *Argumentation in AI*. Springer. 423–441.

[Walton, Reed, and Macagno 2008] Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press. Paperback.