

# Bouillon: reputation-driven universal P2P aggregator

Victor Grishchenko  
Institute of Physics and Applied Mathematics  
Ural State University  
51 Lenina st.  
Yekaterinburg, Russia  
gritzko@ural.ru

## ABSTRACT

The Bouillon project has an objective of building universal P2P aggregation platform employing social links and user reputations to exchange end-user's evaluations for massive parallel collaborative filtering of digital content.

Currently, the project is in stage of public testing, available at <http://bouillon.math.usu.ru>.

## 1. DEMAND

### 1.1 Signal/noise ratio

Spam is the problem that definitively has its roots in the lack of explicit trust model in the current Internet technologies. According to a recent report on blogosphere growth,

These spam pings are fake or bogus notifications that a blog has been updated; in some cases, these *spings* can amount to a denial-of-service attack, and can sometimes account for as much as 60% of the total pings Technorati receives. However, we've built a sophisticated system that mitigates the spings, and helps to keep spam blogs out of our indexes. Beyond that, about 9% of new blogs are spam or machine generated, or are attempts to create link farms or click fraud. [2]

According to MessageLabs statistics [3], 60% to 95% of e-mail messages is spam (depending on period). Another service *spamttest.ru* reports 75-80% of Russian mail traffic to be spam [4]. So most probably, blog spam has some room for growth.

There are also some more sophisticated forms of lowering signal/noise ratio, such as online actors:

Nvidia stands accused of hiring online actors to create dozens of personae in online forums, where they won gamers trust by talking about subjects unrelated to Nvidias products, and then splurged in an orgy of sock-puppet boosterism of Nvidias stuff. [6]

In the Russian segment of the Internet this phenomenon is called The Brigade with the meaning of KGB people toeing the current Party Line on countless forums and sites. Actors having commercial agendas are also widespread.

Copyright is held by the author/owner(s).  
WWW2006, May 22–26, 2006, Edinburgh, UK.

Some self-organized “signal killer” effects such as resident flammers, trolls and newbies are described in [1].

Decay of Usenet (“September that never ended” [9]) was a good example of all those effects combined.

All the phenomena mentioned above have something in common. Namely it is the fact that those information exchange technologies have no tools for explicit expression of trust. This lack creates an opportunity for stealing trust and, generally, for abuse.

One of the Bouillon project objectives is to develop an information exchange environment where every message is marked with trust/reputation/relevance/semantic distance. I.e. to explicitly express trust information. The approach also assumes that for every message there is someone responsible for it; also that the list of responsible entities is manageable (not every IP on the Net).

### 1.2 Duplication of effort

The current trust architecture relies on such indirect methods as domains of physical control (servers, web sites), human control (moderation), automated filters (e.g. Bayesian) etc. While applied to news, for example, the reliance on physical domains of control for trust and filtering creates a stereotypical news propagation method: reprinting (and, sometimes, linking) from site to site. This widespread pattern is successfully exploited by Google News today. Copying news from one domain of control to another is generally unnecessary work, providing significant possibilities for optimization.

Another promising direction is that in the most of current schemes end-user's work of evaluating content is wasted. Some systems, namely Slashdot, have mechanisms for user attention reuse. Sadly, the Slashdot ecosystem [7] is too closed. Feedback forms of “rate and review” type is another attempt to reuse readers' attention. Search engines record user's clicks. Bayesian mail filters let reuse spam recognition effort locally. One possible consequence of explicit expression of trust is direct involvement of user's evaluations (opinions) into message propagation; thus an important message may propagate by itself, *verbatim*, without needless reprinting.

Taking this approach we may explain success of PageRank by its effective reuse of web page editor's efforts making links to worthy resources. However, global ranking of web pages is not a very sharp tool; in particular, it is hardly applicable to resources of local value. P2P propagation is supposed to be much more sensitive and personalized tool, working well for local communities and news.

So, another objective of Bouillon is to reuse efforts of read-

ers including those of local value.

### 1.3 Fragmentation

Another problem with roots in the area of trust is fragmentation. This problem is closely related to duplication of effort. Because of the current trust technology dependence on domains of physical control, knowledge is scattered over numerous such domains, which have no better ways of communication than manual copying and linking. E.g. how can I get the latest news in the area of trust and reputation science? The current method is to track dozens of sites, blogs, boards and mailing lists. Luckily, RSS automates this process, but again, there is little reuse.

Another example of fragmentation is forums or blog comments. The same news subject may be extensively commented on several resources. On every resource, there are *some* worthy comments many people would possibly want to read. A reader has to check every resource, just like hundreds of readers before him, or to give up. While we are using human control over physical domains as a main trust technology, no automated comment filtering and propagation technology is possible.

So, one more objective for Bouillon is to put all the network content into single semantically addressed space.

## 2. PRINCIPLES

Bouillon is a universal P2P aggregator. The only restriction on the aggregated content is that it has to be XML. Bouillon employs social links among participants to propagate interesting items. Thus, it is a social collaborative filtering system. Potentially, a massive distributed social collaborative filtering system.

The vision of Bouillon is to let every user act as an author, critic, editor and reader – and to maximize effort reuse relying on the ubiquitous trust model.

The main idea behind Bouillon is rather simple. Every peer in the network has some acquaintances (friends, IM contacts). There is a number in  $[0.0, 1.0]$  which denotes reputation (similarity) of every friend. Every message is marked by distance from origin. Initial value of distance at the point of origination is 1.0. The message is propagated by hops, from friend to friend. At each hop the message has its distance modified according to the reputation of the hop source peer in the eyes of the hop destination peer,  $distance' = distance \times similarity$ . If a message is positively evaluated by some reader, it starts propagating from a new starting point.

Every participant is ideally able to retrieve all messages on a given topic up to some distance threshold.

Messages are ordered in a tree.

The content carried is arbitrary XML.

### 2.1 Semantic addresses

I use the term “semantic addresses” in the sense of address scheme involving no physical (*w3.org* as a nickname for *128.30.52.46*) or personal (*mailto:gritzko@ural.ru*) details. A sample of a semantic address from the current Bouillon sandbox is *oc:/FAQ/Is the content persistent?.text/Some is persistent.text*. It is not assumed that address has to be memorizable char-by-char.

Today, a typical approach for a site is to use some semantic addressing scheme inside. Usually it is reflected in the right part of an URI *http://www.pathfinder.com/money/-*

*moneydaily/latest/* [5] or *http://economist.com/markets/-indicators/*. Blog categories is another example. Still, semantic addressing schemes are usually local to a particular domain of control. One may think of Bouillon addresses as of “only the right part” which acts as a global address.

Concept of *oc:* semantic addresses slightly differs from e.g. *http:* URIs. Although it is possible to guarantee that a message pointed to by an *oc:* URI is the same for all the readers, it is not guaranteed that every reader in the network is able to obtain that message. It is also not guaranteed that children messages are the same for every reader. And it is mostly guaranteed that expected relevance for the pointed message is different for different readers.

Hyperlinks in the Bouillon network have the same specifics as *oc:* URIs, albeit there are some methods to ensure message is available for those who can reach a link to that message.

Currently, Bouillon is limited to hierarchic categorization. The problem of faceted classification is solved theoretically. Plain tagging is unavailable.

### 2.2 Semantic distance

Semantic (or reputation) distance is the shortest path on the acquaintance (IM contact, FOAF) graph assuming that its directed edges are weighted with respective reputations in  $[0.0, 1.0]$ .

The term semantic distance is used interchangeably with reputation distance. It is reasonable if we talk about the semantic distance between meanings of the same word as different persons understand it. The reputation assumed is also the specific case of opinion similarity reputation. Thus, terms reputational or semantic correspond to two ways of understanding difference in opinions. One may interpret that other peer is lying or just speaks another language (i.e. uses the same word with some meaning different from ours). From the standpoint of Bouillon there are no practical difference here.

### 2.3 Semantic content

Basic requirement for Bouillon network contents is to clearly separate logic from presentation. At least.

Types of content form an OO-like hierarchy (i.e. *urn:oc:text:comment* inherits all the attributes of *urn:oc:text*). At this moment, there are two top-level types: *urn:oc:text* and *urn:oc:taxon*. Text is any textual content. Taxon is ideally any body-less content (e.g. folders) – its only meaning is in its name (“Wittgensteinian” content). It is prescribed that every *urn:oc:ancestor:descendant* has to be processible, to some extent, by *urn:oc:ancestor*’s code.

Note. The reader will likely think that “semantic content” refers to RDF [10]. Although the current version of JBouillon does not work with RDF, it is possible to distribute RDF over Bouillon.

### 2.4 Social networks

Bouillon relies on social networks for trust relationships and information propagation. First, because it is people who do the real work. Second, because the interpersonal form of trust is natural, simple and intuitive. Third, because one may practice Lego approach using basic Bouillon elements (peer and contact) as building blocks for more complicated systems.

### 3. THE BOUILLON PLATFORM

So, what is the Bouillon? This technology is rather universal. E.g. you may think of the Bouillon as of a distributed P2P moderator-less forum, where each user is, in fact, a moderator. Or as of a distributed RSS news aggregator employing reputations. And so on and so forth.

#### 3.1 Protocol

Bouillon protocol is implemented over XMPP IM network (Jabber). This provides Bouillon with the ability of near-realtime message delivery and a ready-made social network of IM contacts. The protocol implements simple request-response message exchange pattern between two adjacent peers. Considering the whole network, message propagation patterns are more complicated. The mainstay of the network architecture is interaction of two types of messages: opinion and opinion request. Opinion is a claim by some *author* that some message is relevant to its parent message to some degree (*relevance, agreement*). The opinion's copy at every peer is always marked with semantic distance from the author to this peer.

Opinion request is a request by a peer to provide him with all the opinions on relevance of children of a given message; those opinions must have semantic distance above some given threshold. This also involves negative opinions.

Thus a peer may obtain opinions on children of some message (e.g. replies or folder contents) by propagating his requests to some neighborhood. By evaluating some message (i.e. by issuing an opinion) a peer may let the message propagate further or may stop it.

For further details on message formats see Sec. 3.3

#### 3.2 Terms

Here is the list of key Bouillon terms:

- author is the original author of a message or an opinion on relevance of a message
- branch is another name for a message, with an accent on its position in the tree
- contacts (friends) are the usual IM contacts (see Fig. 1; contacts of Alice are Frank, Bob, Eugene, Hector)
- reputation (of a contact) is a value in  $[0.0; 1.0]$  which is derived from experimental comparison of the contact's (Bob's) prior opinions (advices) and the peer's (Alice's) posterior opinions (evaluations). 1.0 mean perfect match; 0.0 is a perfect disagreement. Reputation is always personalized, e.g. the reputation of Bob in the eyes of Alice  $r_A(B)$ . No global reputation values exist in the system. This understanding of reputation is a generalization of a metric proposed in [8].
- reputation distance is a cumulative reputation value for the shortest chain from one peer to another,  $r_A(C) = r_A(B) \times r_B(C)$ . Every message propagating in the Bouillon network is weighted with reputation distance of the path it traveled. Reputation distance is not symmetric.
- distance threshold is the maximum reputation distance the requesting party could tolerate. Opinions from beyond the threshold are of no interest. Distance threshold of your requests could be changed using the ?-slider

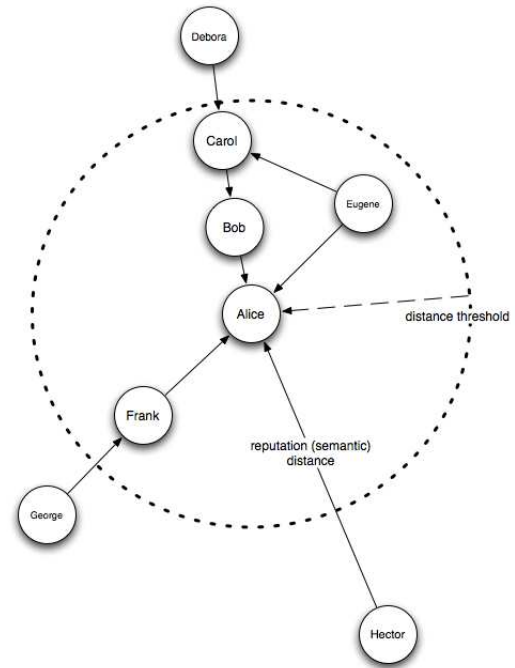


Figure 1: Alice-Bob-Carol example.

(Fig. 3). See Fig. 1 where the dashed line depicts reputation distance threshold set by Alice; opinions by Frank, Bob, Carol and Eugene are accepted; Hector is beyond the threshold although he is an immediate contact of Alice. Probably, he was too abusive and thus got low reputation. Or, his opinion is of no value on the subject considered. (JBouillon employs subject-dependent reputation values.)

- agreement is a value in  $[0.0; 1.0]$  which characterizes relevance of a message to its parent (e.g. relevance of a reply to the original message in some forum). Agreement is set by a user with the !-slider. (And thus an evaluation is made and an opinion is issued.) (expected) relevance of a message is our best educated guess: agreement of some peer on that message times reputation distance to that peer. In the case of multiple available opinions a simple weighting formula is used. Relevance threshold is the actual parameter set by the ?-slider.

#### 3.3 Messages

Every message has the following attributes:

- author
- name/URI/branch
- contact (the message came from)
- distance from the author to the current peer
- distance from the current peer to the author
- TTL

```

<?xml version="1.0" encoding="UTF-8"?>
<post
author="Victor Grishchenko"
branch="oc:///FAQ/Is+the+content+persistent?.text/
Some+is+persistent.text"
distance="1.0"
rdistance="1.0"
timercvd="1139234487450333000"
ttl="9223372036854775807"
xmlns="urn:oc:token">
<oct:text
xmlns:oct="urn:oc:text"
oct:date="2006.02.06 19:01:27 +0500">
Your messages and messages you
evaluated are stored locally.
Transit content is not stored.
</oct:text>
</post>

```

Figure 2: A sample post (oc message containing a simple text message).

### 3.3.1 Post

Post carries an actual message body (XML). Every peer permanently stores bodies of all the evaluated messages. Considering XML formats used, there is a requirements of inheritance dictated by Sec. 2.3. In the current implementation, message hierarchy is implemented by *jbouillon.oc* Java class hierarchy (e.g. *jbouillon.oc.text.comment.Post.class*).

See Fig. 2 for an example of plain text message.

### 3.3.2 Post Request

No special attributes. Post requests are propagated by footsteps of respective opinions to reach the opinion's author (i.e. the recommender).

### 3.3.3 Opinion

Opinion has a special attribute of **agreement**, i.e. the degree of relevance of a message to its parent message in the eyes of the opinion's author.

### 3.3.4 Opinion Request

Opinion request has a special attribute of distance threshold which is equal to relevance threshold at the point of origination. Opinion requests flood network in all the promising directions (see Sec. 3.4, the climbing process) as resource constraints permit.

## 3.4 Consistency principles

There are two consistence principles that let Bouillon network work.

PRINCIPLE 1 (? CONSISTENCY). *Child opinion request has the threshold higher-or-equal than the parent's.*

PRINCIPLE 2 (! CONSISTENCY). *Child opinion semantic distance is equal-or-lower than the parent's.*

That is, if a peer has evaluated a message, all unevaluated ancestors are also evaluated automatically. To issue a request on a child message contents, a peer has to issue equal-threshold request on the parent message contents.

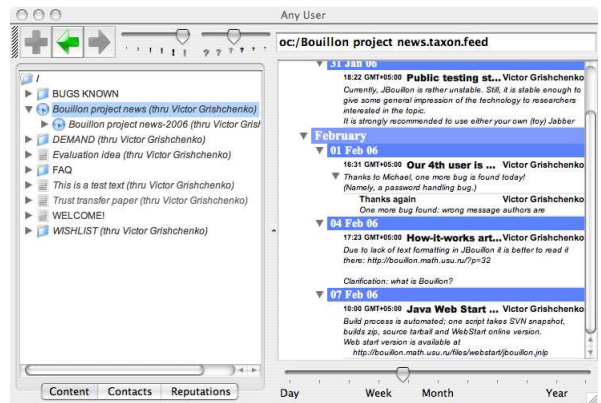


Figure 3: JBouillon GUI

These principles underlay the *oc* stack which consists of four layers: storage, matching, consistency and user layer. Peer-to-peer communication proceeds at the matching layer. All layer-to-layer and peer-to-peer connections use the same asynchronous interface. Although this stack seem to be the main contribution of the Bouillon project, it is not discussed in detail here.

These principles enable the algorithm of focused search (climbing). Having opinions (replies) for message *a/b*, a peer knows which contacts may deliver opinions on *a/b/c* satisfying semantic distance threshold requirement. Thus the search process is directed to promising peers without disturbing those peers who surely can't deliver a response. Although this may seem to be just an optimization, this is the actual enabler of the technology. Tree-climbing algorithm differentiates Bouillon from query flood or DHT solutions.

## 3.5 GUI description

JBouillon client graphical user interface (Fig. 3) has the following Bouillon-specific components:

- navigation tree
- + button to add new messages
- !-slider to evaluate messages
- ?-slider to set relevance threshold for the current branch (i.e. to get more or less messages)

The particular design of the interface is work-in-progress.

## 4. CONCLUSION

### 4.1 Other relevant research

Today, reputation/trust research is a subject of extensive research by itself. Even more, it seems to me that a good overview of that research-on-research is a thing people may need today, because all kinds of mentioned research are subject both to exponential growth [11] and to considerations of Sec. 1.

Anyway, Bouillon metric has *some* similarity with all the trust/reputation metrics using multiplication as a concatenation function and maximum as an aggregation function in terms of path algebra on social (trust) graphs [12], although

Bouillon model has more dimensions. Orthogonal to the social topology made of participants (peers) there is also the topology of messages (currently, a tree). The actual Bouillon network is kind of a fuzzy subset of a Cartesian product of these topologies.

## 4.2 Current state of the project

At this moment, a limited prototype of the Bouillon platform (JBouillon) undergoes public testing. All the information is available at <http://bouillon.math.usu.ru>.

## 4.3 Future work

One of possible Bouillon extensions is to replace a tree of messages with a directed weighted graph, kind of a wordnet. This approach will let to blend plain tagging and hierarchic categorization into a single technology. A preliminary result is that the hybrid approach has nearly the same order of computational complexity, although it complicates understanding of the technology a lot.

Another direction is to use faceted classification and navigation. This generalization is theoretically solved and is implementable as a trivial ad-hoc over the existing Bouillon platform.

## 5. REFERENCES

- [1] "A Group Is Its Own Worst Enemy", online essay by Clay Shirky, available at [http://www.shirky.com/writings/group\\_enemy.html](http://www.shirky.com/writings/group_enemy.html)
- [2] David Sifry "State of the Blogosphere, February 2006", available at <http://www.sifry.com/alerts/archives/000419.html>
- [3] MessageLabs spam intercepts statistics, available at <http://www.messagelabs.com>
- [4] "Spam 2004" report by *spamtest.ru* (c) <http://www.ashmanov.com>
- [5] "Cool URIs don't change" URI style guide available at <http://www.w3.org/Provider/Style/URI>
- [6] "Did Nvidia Hire Online Actors to Promote Their Products?" available at <http://www.consumerist.com/consumer/evil/did-nvidia-hire-online-actors-to-promote-their-products-152874.php>
- [7] "Slashdot Moderation" available at <http://slashdot.org/moderation.shtml>
- [8] Victor S. Grishchenko: "Redefining Web-of-Trust: reputation, recommendations, responsibility and trust among peers", FOAF'04, available at [http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/jp/redefining\\_web\\_of\\_trust/](http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/jp/redefining_web_of_trust/)
- [9] Wikipedia article on Eternal september: [http://en.wikipedia.org/wiki/Eternal\\_september](http://en.wikipedia.org/wiki/Eternal_september)
- [10] Resource Description Framework (RDF), <http://www.w3.org/RDF/>
- [11] Andrew M. Odlyzko: Tragic loss or good riddance?, <http://www.dtc.umn.edu/~odlyzko/doc/tragic.loss.txt>
- [12] M. Richardson, R. Agrawal, P. Domingos: "Trust Management for the Semantic Web", in Proc. of 2nd Int Sem. Web Conf, 2003