Exploration of Quantitative Factors Affecting the Popularity of Users in an Online Community

Matti Eteläperä[1, 2], Mourad Oussalah[1]
1: CMVS, Faculty of Information Technology, University of Oulu,
Oulu, Finland
2: Pepron Software Services, Oulu, Finland
matson@iki.fi , Mourad.Oussalah@oulu.fi

Ab*stract* — Online communities for discussing different topics have been around since the dawn of the Internet. Whether an individual user is popular or not in the community is highly subjective and difficult to measure with a quantitative approach. In this paper we explore factors and usage patterns on a Finnish football related online community FutisForum2 for receiving votes in a yearly voting for Forum User of the Year. Several quantitative factors are identified in order to calculate correlations between each factor and the yearly voting results. These factors include yearly message network centrality figures, number of message, user quote amounts, number of characters in messages, etc. Although message amounts clearly correlate with the voting results, the strongest correlation was noticed when comparing eigenvector centrality to the voting results. The main outcome of the project was also generation of a database of 11 million messages for further research

***Keywords—NLP,* social network analysis, popularity metrics**

I. INTRODUCTION

The proliferation of Web 2.0 technologies has triggered the rapid development of online communities where users come together around a shared purpose, interest or goal [1], often interacting with each other to exchange ideas, discuss topics or seek special requests. Notably, Facebook, the largest global online community, holds 2.4 billion monthly active users around the world as of the 2nd quarter of 2019. Besides thousands of virtual communities are created daily, sometimes, as part of professional corpora activities, and sometimes, as part of common values shared by the group members. Acknowledging the great potential of online communities in reaching a wider audience and new market opportunities, venture capital groups invested and are still ready to invest billions of dollars in gaining control of active online communities.

On the other hand, many online communities died shortly after their creation. Indeed, intuitively, individual community member's behavior is not only influenced by his/her inherent or cognitive motivations but also by other group member attitudes and activities as well as by the community behavior as a whole. This triggers the importance of identifying key factors that influence the growth and sustainability of the underlined online community. For this purpose, various theories have attempted to contribute to the ongoing research of comprehending users' participation and behavior and reinforce community cohesion /growth. This includes social cognitive theory, social capital theory [2,3], trust theory [4] and social network theory [5]. This paper develops a social network-based approach to comprehend the key factors influencing users' participation in a popular sport-related Finnish online community Futisforum2. The latter hosts over 11 million messages and more than 40,000 users since 2006 about football-related discussion topics. Actors in this field are constituted of football players, their family members, managers, journalists, fan and followers, which often have a mobile life trajectory, hybrid social cultural practices and sociolinguistic repertoires [6]. This makes the idea of exploring the underlined social network appealing.

Social network theory uses methods of depicting and analyzing networks of people to help understand and communicate how they are connected [7,8].

Futisforum2 is the most popular football related discussion forum in Finland, hosting over 11 million messages since its start in 2006. Up to the year 2019, over 66 000 users have contributed to writing up to 220 000 discussion threads. The discussion forum runs on Simple Machines Forum (SMF) software and some extra statistics of the user base are available by the forum. For example, the gender ratio on the forum is around 1 to 16 male versus female and the average age of a user is around 30 years.

Due to massive data scale and increasing noise ratio made by random users, untrusted profiles or simple bots, the popularity of an individual user in an online discussion forum is difficult to evaluate using a pure automated and quantitative approach. In this context, Van Leeuwen [9] argued that authenticity is in crisis because of the multiplicity of factors in late modernity age, which include complexity of membership to community, increasing use of dialect styles, development of personal narratives.

In this paper, we utilize the ground truth information of user's popularity, available by the virtue of yearly voting results in the Forum User of the Year election on Futisforum2 in order to tackle the influence of other factors. The aim of this paper is a) to identify the quantitative factors with the strongest correlation for receiving votes and b) discussing possibilities for predicting voting popularity solely on quantitative behavior on the forum.

Throughout this paper, a new methodology for gathering source data, filtering this data, identifying quantitative factors of interests will be lied down. In the result section, we show how these factors correlate with voting results using visualizations. The following discussion section wraps up the findings and proposes possibilities for estimating the voting behavior. Future work section lays out the conclusion and possible research ideas for further exploration.

II. METHODOLOGY

A. Overall methodology

A general process diagram regarding the different steps for the research work is shown in Figure 1. The process starts with data scrapping from the website, distinguishing basic metadata associated with each user. Next, key statistical data are gathered regarding individual participants test correlation results with yearly user's voting results. Especially, various network centrality measures and influence evaluation will be evaluated and quantified. Finally, some visualization toolkits are investigated to comprehend the potential contributions of individual factors and interact with the analyst, which would ultimately pave the way for future and subsequent analysis.

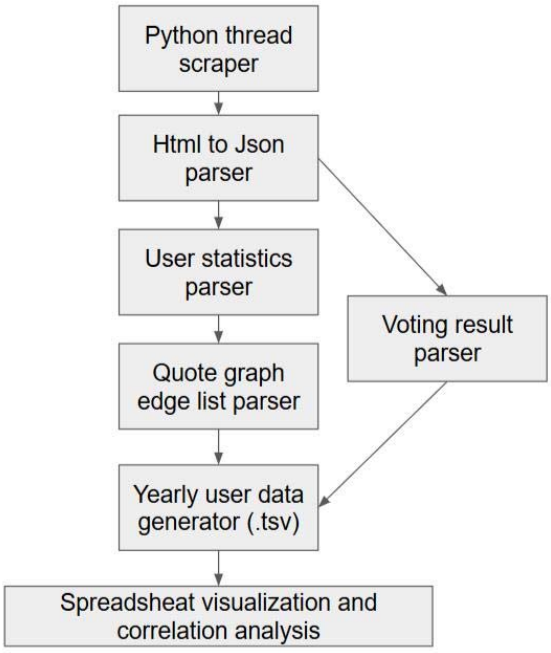


Figure 1. Overall approach.

B. Data collection

As no publicly available dataset exists for FutisForum2, we employed a simple HTTP protocol to scrape all users' messages. SMF features a print mode for discussion threads, in which all messages of a single discussion thread can be downloaded with a single HTTP GET call. An example of this data as viewed through a web browser is presented in Appendix A.

FutisForum2 also features a system in which a registered user can access more topics than an anonymous user, so a new user account was created for accessing this data. This scrapping data is performed using a Python script for downloading all existing messages composed on the website since 2006. The script performs a two-step login process utilizing cookies to access the forum.

C. Identification of quantitative factors for correlation analysis

Several quantitative factors from the source data were identified as candidates for further exploration and subsequent analysis. The data was separated per user and per year (from 2006 to 2019). The following factors were evaluated.

1) *# of messages composed*
2) *# of threads started*
3) *# of times user has quoted another user*
4) *# of times another user has quoted the user*
5) *Average message length*
6) *Degree centrality of user in message network*
7) *Eigenvector centrality of user in message network*
8) *Ratio between user quoted others vs others quoted user*

D. Centrality measure extraction

Factors 2-5 from subsection *C* were extracted from the message threads directly using a Python parser. Factors 1 and 8 were estimated in the spreadsheet phase from this data. While NetworkX library [10] in Python was used to gather Factors 6 and 7. This required yearly data of messages written on the forum, resulting in a directed, weighted graph. The nodes of this graph are users and quotes between users correspond to edges in the network. Therefore, using NetworkX, it was easy to perform various centrality measures for users; namely, degree centrality and eigenvector centrality.

In other words, degree and eigenvector centrality values associated with yearly message thread were constructed using NetworkX commands as exemplified in Figure 2. An obfuscated string "½£{¤" was used as a delimiter in the edge list in order to avoid whitespace collisions as would have happened by using a default network call read_edgelist ().

```
G = nx.read_edgelist(EL_file,
delimiter="½£{¤",data=(('weight',float),))

nx_analyysi[year]["deg_centrality"] =
nx.degree_centrality(G)

nx_analyysi[year]["eigenv_centrality"] =
nx.eigenvector_centrality(G)
```

Figure 2. NetworkX commands to create centrality factors.

E. Preprocessing of data and research database models

JSON and text files were used for storing results from each individual step in the process. As all source data was in HTML form, the preprocessing task was performed by utilizing the Unix stream parsers (sed, jq) and BeautifulSoup HTML parser for Python. The following example illustrates the information stored in the files for user statistics in JSON format:

{
YEAR:

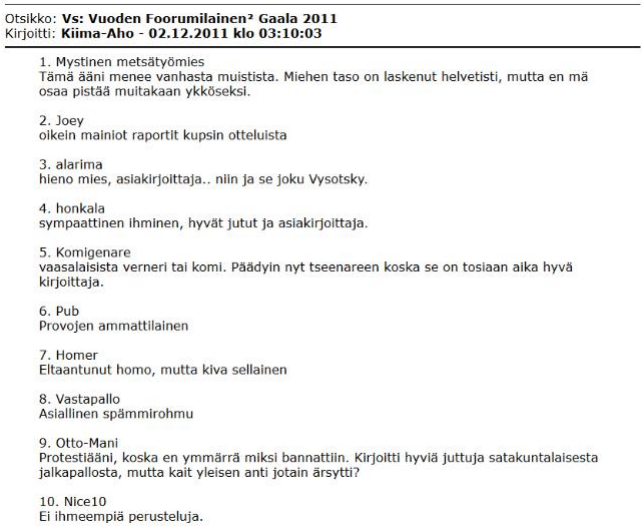
```
{
USERNAME: {
    "Msg_written": int,
    "User_quotes_other": int,
    "Other_quotes_user": int,
"Eigenvector_centrality": float,
    "Degree_centrality": float,
    "Characters _written": int,
    "Voting_points": int
        },
    USERNAME_2..
    }
YEAR_2…
}
```

F. Extraction of voting data

Yearly voting results from 2007 to 2018 were extracted from yearly voting threads. A Python parser was written to perform this task and the results are stored in yearly files. All usernames were converted to lowercase to minimize the number of errors induced by possible typos made by voters.

The parsers written for this task attempts to filter out invalid votes by the means of string matching. Voting data from the year 2007 was omitted in total due to low-quality inputs leading to parsing difficulties.

An example of a single valid vote is depicted in Figure 3.


Otsikko: **Vs: Vuoden Foorumilainen² Gaala 2011**
Kirjoitti: **Kiima-Aho - 02.12.2011 klo 03:10:03**

1. Mystinen metsätyömies
Tämä ääni menee vanhasta muistista. Miehen taso on laskenut helvetisti, mutta en mä osaa pistää muitakaan ykköseksi.

2. Joey
oikein mainiot raportit kupsin otteluista

3. alarima
hieno mies, asiakirjoittaja.. niin ja se joku Vysotsky.

4. honkala
sympaattinen ihminen, hyvät jutut ja asiakirjoittaja.

5. Komigenare
vaasalaisista verneri tai komi. Päädyin nyt tseenareen koska se on tosiaan aika hyvä kirjoittaja.

6. Pub
Provojen ammattilainen

7. Homer
Eltaantunut homo, mutta kiva sellainen

8. Vastapallo
Asiallinen spämmirohmu

9. Otto-Mani
Protestiääni, koska en ymmärrä miksi bannattiin. Kirjoitti hyviä juttuja satakuntalaisesta jalkapallosta, mutta kait yleisen anti jotain ärsytti?

10. Nice10
Ei ihmeempiä perusteluja.

Figure 3. Typical valid voting result before parsing.

In this example, the parser creates a Python dictionary with yearly and voter data included. The list of votes is ordered and the parser later gives points 1 to 10 depending on their order.

```
{"2011":
 {"Kiima-Aho":
[       "Mystinen metsätyömies",
        "Joey",
        "alarima",
        "honkala",
        "Komigenare",
        "Pub",
        "Homer",
        "Vastapallo",
        "Otto-Mani",
        "Nice10"
        ]
    }
}
```

This enables us to generate a weighted and a directed voting network. In this respect, the scores of individual user votes are summed up so that yearly results are available for a fast query tasks.

G. Correlation analysis and visualization

Correlation analysis between yearly voting results for users and the factors under examination was performed. The methodology was to first select users who received any votes and then compare this set to the selected factors as pointed out in subsection C. This leads to a scatterplot to which a linear trendline is drawn to see the direction of the correlation.

To facilitate possible future work with the data by other researchers, the source codes used for both pre-processing and data analysis are made public in GitHub. Data scraping scripts are not supplied in order not to overwhelm the FutisForum2 databases, but raw source data used in this work is available from the author directly.

III. RESULTS

To comprehend the effect of various social network-based factors, we focus on those users for which the number of votes gathered is known. Therefore, we explore the various factors pointed out in subsection C of the previous section to assess their correlation with the number of voting points. This enables us, for instance, to evaluate the extent to which the degree centrality of the user predicts the number of votes gathered. A general statistical description of the data is presented in Table 2.

Table 2. Overall statistics on the employed data

# Users (U) with voting points, yearly average	566
Total number of users between 2008-2018	62594
Messages per U (sample year 2015) (max, avg, min)	(7652, 570.8, 1)
# user has quoted others in 2015 (max, avg, min, std)	(3348, 266, 0, 397,7)
# other have quoted user in 2015 (max, avg, min, std)	(4149, 250, 0, 366,9)

For illustration purposes, we initially present results of scatterplots from a randomly selected year (here 2015). Figures 4-11 illustrate the behavior of the various factors with respect to voting points.

Figure 4 highlights the scatterplot of new threats initiated with respect to gathered voting points.

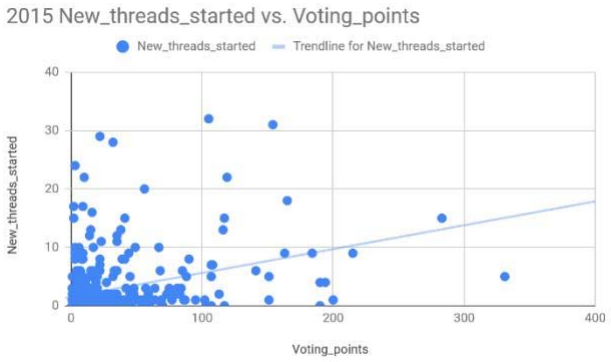


Figure 4. 2015, Threads started vs voting points.

Whereas the influence of the number of written messages with respect to voting points is shown in Figure 5.

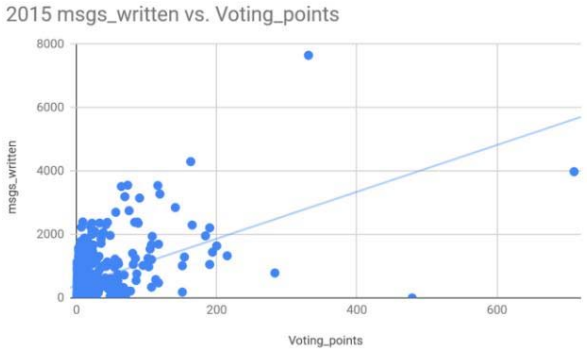


Figure 5. 2015, Messages written vs voting points.

Figures 6 and 7 highlight the influence of the user-degree centrality and Eigen-vector centrality, respectively, on the number of votes collected.

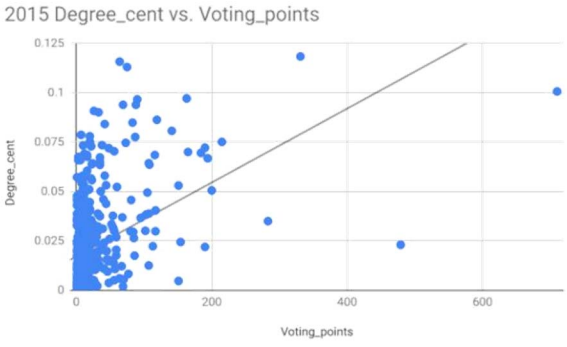


Figure 6. 2015, user degree centrality vs. Voting points.

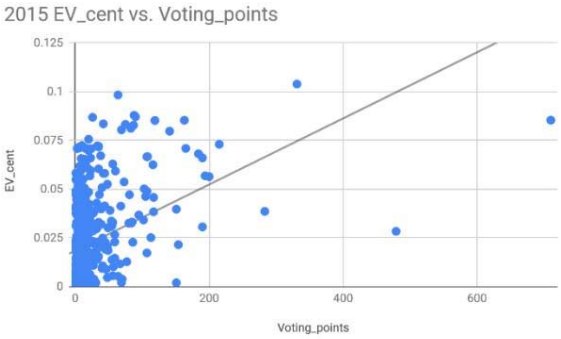


Figure 7. 2015, Eigenvector centrality vs Voting points.

Figure 8 quantifies the relevance of user-quotation; speculating that the mentioning of a User A by the given User B can be an indication that A will gather higher number of votes. For this purpose, we consider the ratio of the number of quotes to the total number of user messages.

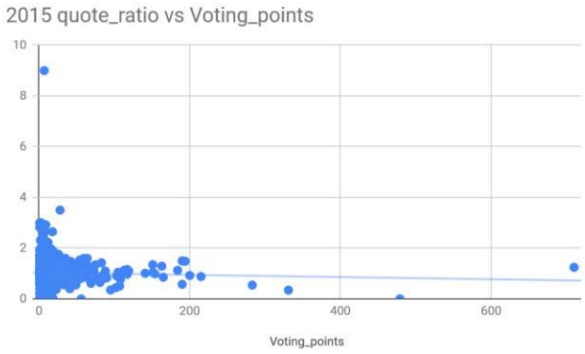


Figure 8. 2015, Percentage of user messages including a quote of other users vs Voting results.

Instead of voting point, Figures 9 and 10 highlight the influence of user quotation and other quotations, respectively, with respect to number of messages written by the user.

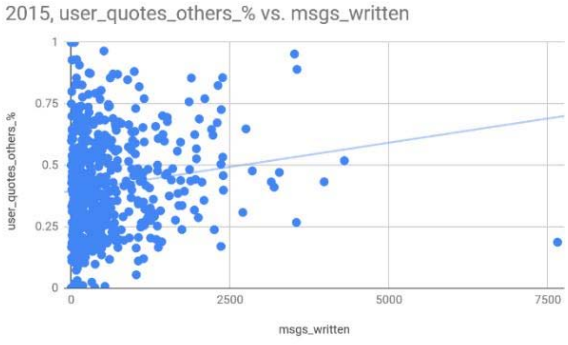


Figure 9. 2015 Ratio between user messages vs. other users quoting user

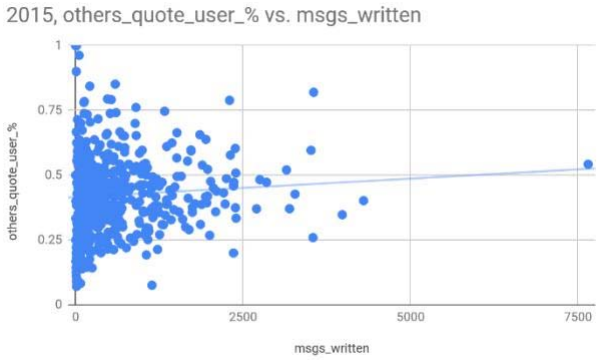


Figure 10. 2015 Other quotations versus number of messages written.

Finally, results in terms of average message length is shown in Figure 11.

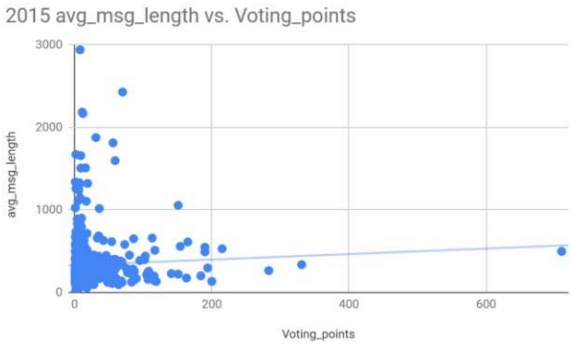


Figure 11. 2015, average message length vs. Voting points

A quick reading of the results highlighted in the figures 4-11 indicates at least two milestone results. First, the distribution of voting points with respect to individual factors is highly non-uniform where the quasi-majority of voters get held less than 200 voting points, and only very few users exceed 200 voting points. Second, there is no individual factor that shows the potential to predict user's voting points. Therefore, the idea of universally accepted factor should be omitted. In order to provide a more quantitative evaluation of individual factors, a Spearman's rank correlation was run to determine the relationship between n=546 voting results and seven factors. Spearman's rank correlation was chosen as it is not sensitive to highly non-gaussian datasets present for example in this study. The results are shown in Table 1.

Table 1. Evaluation of correlation strength of factors vs voting results.n=546.

Factor	Spearman's rho	Significance t-test
Messages written	0.453	0
Eigenvector centrality	0.383	0
Degree centrality	0.402	0
Message topics started	0.341	0
Average message length	0.114	0.007
User quotes others ratio	0.115	0.007
Others quotes user ratio	0.05	0.247
Quote ratio (user quotes others vs. others quote user)	0.071	0.099

Discussions

The correlation analysis reveals the relevance of message amounts, discussion thread starts, and both degree and eigenvector centrality figures to the voting results of the specific year 2015 under examination. These results were pretty expected in many scenarios. For instance, the steady stream of messages on the discussion board serves as a way to keep one's name relevant in the discussions. Also the fact that the most popular users were highly networked and regularly discuss with other popular people was to be expected.

What was more surprising was the very limited effect of message lengths and how much both the user quotes other people and how much other people quote the user on the voting performance. The Spearman's rank correlation number of these factors was low and the typical statistical significance test for $p<0.05$ failed for these two last factors as highlighted in Table 2. Nevertheless, this study can be extended from various perspectives. First, the use of more elaborated multivariate statistical based approach to comprehend the voting behavior seems appealing. Second, the temporal aspect in the correlation analysis has not been taken into account in this study, although, we believe that such information could shed the light on potentially more relevant linguistic and/or network related features. Third, this work attempted to find correlations between users receiving at least one point in the voting and the factors chosen. However, more useful information could be gathered from the vast mass of users who did not receive any votes at a given year. This information has so far been neglected in the current study. Fourth, the current analysis has also opened up the interest for new potential factors/features. This includes number of emojis used, number of internet link posted and number of news posted. Fifth, it could also be interesting to include the sentiment score for the messages written by the user. This could be implemented for Finnish language by stemming the messages of users and performing a sentiment

vocabulary match on each message. The resulting score would be a similar factor to the factors explored in this paper.

CONCLUSION AND IDEAS FOR FUTURE WORK

The main takeaway of the work is that most quantitative factors selected for the analysis behave reasonably. The more messages a user composes, the more likely he or she is to get points in the yearly voting.

It seems that the variance between individual factors versus the number of points received in voting is very high. Because of this, it can be argued that this approach alone does not provide enough information, for example, to estimate the yearly voting results beforehand. A more advanced, multi-variate statistical scheme would have to be used to get correlations with less variance. Despite this, the results do show that both network centrality factors (eigenvector and degree) seem to have the strongest correlation on the voting results of all factors explored. So, although this work gave many meaningful insights into the voting data and factors affecting voting results, there is room for finer-grained work.

As a side effect of this work, a huge database of well-formed data about a medium-sized, topic restricted Internet forum in Finland was created for further research.

REFERENCES

[1] Figallo, C. *Internet World: Hosting Web Communities.* John Wiley & Sons, Inc., New York, 1988.

[2] C.-M. Chiu, M.-H. Hsu, and E. T. G. Wang, "Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories," *Decis. Support Syst.*, vol. 42, pp. 1872–1888, 2006.

[3]. Preece, J. Supporting community and building social capital. *Commun. ACM 45,* 4 (Apr. 2002), 37–39.

[4] Garbarino, E., and Johnson, M.S. The different roles of satisfaction, trust, and commitment on customer relationships. *Journal of Marketing, 63,* 2 (1999), 70–87.

[5] Butler, B. Membership size, communication activity, and sustainability: The internal dynamics of networked social structures. *Information Systems Research 12,* 4 (Dec. 2001), 346–362.

[6] Blommaert, J. The sociolinguistics of globalization. Cambridge: Cambridge University Press, 2010

[7]. Andrews, D. Audience-specific online community design. *Commun. ACM 45,* 4 (Apr. 2002), 64–68.

[8]. Kim, A. *Community Building on the Web.* Peachpit Press, Berkeley, CA, 2000.

[9] Van Leeuwen, T. What is authenticity? Discourse studies 3 (4), 392-396., 2001

[10] NetworkX, https://networkx.github.io/, visited 13.5.2019

Appendix. Example of Dataset

Otsikko: **Su 5.10 klo 18.30 AC Oulu - Atlantis FC - Syysillan huumaa**
Kirjoitti: **Sakallio - 04.10.2008 klo 12:43:52**

(http://www.on24.fi/images/c/6017_c_18397.jpg)

Otsikko: **Vs: Su 5.10 klo 18.30 AC Oulu - Atlantis FC - Syysillan huumaa**
Kirjoitti: **Brindellone - 04.10.2008 klo 13:06:37**

Henri Hanhela saa avauspaikan ja iskee kaksi maalia Atlantiksen verkkoon 3-0 kotivoittoon päättyvässä ottelussa. :tuoppi: Tuntui ihanalta lukea viimeisin uutinen ACO:n sivuilta kun kaikki paineet lysähtivät poies:

Lainaus
Tappion myötä AC Oululla ei käytännössä ole enää mahdollisuuksia nousta kahden parhaan joukkoon, joten puheet liiganoususta voidaan siirtää vuodella eteenpäin.

Loppu ne Malisen hopotukset. Nyt jätkät rennosti piiiiiitkästä aikaa kotivoittoon. :tuoppi:

Otsikko: **Vs: Su 5.10 klo 18.30 AC Oulu - Atlantis FC - Syysillan huumaa**
Kirjoitti: **Ismona - 04.10.2008 klo 16:49:29**

Oulu, viekää!

Otsikko: **Vs: Su 5.10 klo 18.30 AC Oulu - Atlantis FC - Syysillan huumaa**
Kirjoitti: **gulp - 04.10.2008 klo 20:02:31**

En jaksa vaivautua paikalle >:(, kausari jakoon.

Otsikko: **Vs: Su 5.10 klo 18.30 AC Oulu - Atlantis FC - Syysillan huumaa**
Kirjoitti: **Riquelme - 05.10.2008 klo 12:09:11**