Slide 1

My name is Bruno Leite Alves and I will talk about the role of research leaders on the evolution of scientific communities.

This work is done together Fabrício Benevenuto and Alberto Laender.

We are from Federal University of Minas Gerais, from Belo Horizonte, Brazil.

Slide 2

The society is organized in communities and there are a lot of kinds of communities, such as: fans of a sport, fan of a celebrity, friends communities, scientific communities and so on.

We may name these communities as social networks, and inside of these social networks there are individuals that influence and are influenced by others.

Inside of these communities, there are opinion leaders which are people able to affect the opinion of other people. In this way, groups of leaders or influential are able to affect the dynamics of the entire communities.

Slide 3

In this context, we have the scientific communities which are a specific type of social network, where the researchers share knowledge, make connection with other researchers, join in scientific groups and publish research results.

Slide 4

Here, our goals is study the dynamics of scientific communities and identify leaderships in this communities, here we named community core. We are also interested in the properties of the community core.

Slide 5

So, the rest of the talk is organized as follows.

Next, I am going to describe the data we used to study scientific communities.

Then, we describe how we extract the core of a community, and finally we present a series of analysis about the communities core.

Slide 6

We used the DBLP data, a digital library with 2.2 million publications from 1.2 million authors. How the DBLP has a lot of information, for a better analysis, we selected 24 flagship conferences of major ACM SIGs and we considered each conference as a scientific community.

Slide 7

So, here it is possible to observer the 24 SIGs selected, all of our analysis is based in these conferences.

Slide 8

In this context, where we would like to identify the members most important, we needed quantify the importance of the researchers inside of communities.

So we define this metric: Core Score. The core score estimates a researcher's importance within a community and the core score of a researcher r into a community c in a specific time t is given by h-index of a researcher times the number of publications of a researcher into a community in a specific time.

An important piece of this metric is the h-index, so we needed obtain this value.

Slide 9

So, we tried using the Google Scholar, a tool important which is widely used among scientific communities to calculate the h-index. For it, we randomly selected 10 researchers for each conference, but only 30% of DBLP authors had a profile at Google Scholar. It happened because in this tool, the researcher need join and create his profile himself.

Slide 10

In this hand, we needed an alternative to obtain the h-index, so we chose to use the shine project, a tool which makes the estimation of h-index of conferences. They use the number of references of the papers to estimate the h-index, so we match this information with DBLP and we inferring the h-index of the researchers.

Slide 11

To validate this estimate, we calculated the Pearson correlation coefficient. It is possible note in the graphic and the value of Pearson Coefficient a strong positive correlation between the inferred value and Google Scholar value.

Slide 12

After did it, we needed to define two important thresholds to extract the core community.

The first one is the size of core community, one time we have a ranking based in the core score.

The second one is the temporal sliding window size to analyze the core community over the time.

We used the resemblance and angular coefficient to help us. The resemblance is a measure which identifies the change in the network over the time and angular coefficient it is responsible to identify the inclination of the lines.

Slide 13

After the calculations we defined the core community size as 10% and the temporal sliding window size as 3 years. Here it is possible observe the SIGMOD conference, but the other conferences follow the same behavior.

Slide 14

We would like to validate the ranking of the core score, so we selected two important researchers of scientific communities, the first one is Jon Kleinberg, the keynote speaker of this conference this year. Here, we may note the core score of Jon Kleinberg's communities.

Slide 15

According with the threshold defined as 10%, we may observer Jon Kleinberg as part of the community core of important conferences, as STOC, KDD, SIGMOD and so on.

Slide 16

Other important researcher selected is Luis von Ahn, he is also a keynote speaker of the www conference this year. It is true that he has a difference between Jon Kleinberg, because Luis dedicated in a specific group of conferences.

Slide 17

We may see here that Luis became part of community core of CHI and he was close to became part of the community core of conferences as SIGCSE and DAC.

Slide 18

A second validate we proposed is identify conference which has award for the best researcher and observe whether this information match with community core.

It is possible note a strong correlation between this two information, only the SIGCOMM has a small value, it is because we selected just the flagships conference and this SIG has a lot of important conferences, so the awarded researchers published their works in others conferences.

Slide 19

Our analyses are performed under two perspectives. The first consists on analyzing the network evolution year by year by accumulating nodes and edges to a single final snapshot of the graph. We may observer the size of largest connected component, in generally, grow up over the time.

Slide 20

The second perspective consists of analyzing snapshots constructed based on nodes and edges created on a predefined time window. Here, using the sliding window, we can note the variation over years. This motivated us to investigate the role of core community in this case.

Slide 21

The focus of this work is understand the effects of the community core in the properties of the networks, so, using the sliding windows, we can note that average degree is higher for core members than non-members.

Slide 22

Differently of the average degree, the clustering coefficient is smaller in community core, so the core might act like hubs in the network, by connecting different groups with small intersection.

Slide 23

So, to identify if the members of community core are hubs, we calculated the betweenness of the community core. The betweenness show which a higher number of shortest paths include the core. It confirms that members of community core are hubs.

Slide 24

To understand how the core score evolve over the time, we propose this metric, the average core score. In the old conferences, it is possible note a similar behavior, where the average core score increase over the time and after a time, it starts to decrease.

Slide 25

And for early conferences, the average core score has a behavior more instable. We can speculate innumerous factors for this behavior:

- expansion or reduction in the number of published papers

- raise and fall of hot topics with ability to attract

- or loose important core members

- members involved in the conference organization

- and so on.

However, disregarding what caused these variations, we want to investigate if such variations can directly impact the network structure.

Slide 26

In this way, we calculated the correlation between the average core score and the other complex networks metrics.

Except for clustering coefficient, every metric have a strong correlation with average core score, in other words, if we increase or decrease the average core score, this metrics also change.

Slide 27

Well, with this work, we conclusion which:

The core community work as bridges that connect smaller clustered research groups, and finally, the core community tends to increase the average degree of the network and decrease the assortativeness .

Slide 28

As future work we propose make analysis of other kinds of network such as massive multiplayer games and on-line social networks, and calculate the h-index according to the time.

Slide 29

Questions?

Slide 30

Thank you very much!