Sourav Bhowmick Speaks Out on Drawing Your Queries, Ethics, and Drawing for Ethics

Marianne Winslett and Vanessa Braganholo

Sourav Bhowmick

https://personal.ntu.edu.sg/assourav/

Welcome to ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today I have here with me Sourav Bhowmick, who is a professor at Nanyang Technological University in Singapore, better known as NTU. Sourav is an ACM Distinguished Member. He received the VLDB Service Award, and a Lecturer of the Year Award from NTU. Sourav is also a member of the Initiative on Diversity and Inclusion in Database Conference Venues. His PhD is from NTU. So, Sourav, welcome!

Thank you very much for inviting me, Marianne. I'm really honored to be in this set of interviews.

You've worked in several classical areas of data research, but also in an unusual one: human data interaction. Can you tell me about that?

I explore the interaction of users and data through a visual interface. We database folks love query languages. We have invented query languages for all sorts of different data types – we started with SQL and now we have graph query languages. But most domain experts don't use these languages. Many do not want to write any queries using query languages: they prefer to draw or to have a very intuitive natural language-based interaction. That's why visual interfaces for queries are very powerful for many domain experts.

Visual query interfaces offer a lot of interesting research opportunities which we haven't explored in the past. For example, twenty years ago I was working on XML because a lot of biological data was in XML format. If the user queries this XML data by drawing a tree structure, we get the benefits of the query being revealed incrementally to the query engine, because the user is drawing the query incrementally. We also get the benefit of the latency of a human drawing on an interface, and the research community had never explored how to leverage that. Our paradigm was that the query engine was not going to do anything until it got the complete query, and then it would find the best way of processing the query.

In a scenario where the queries are getting expressed incrementally as the user draws them, we have the opportunity to partially process these queries while the user is drawing, by exploiting the latency of the human interaction with the user interface. This allows us to integrate query processing and query formulation much more tightly than was possible before, opening up new opportunities such as efficient query feedback and query response time. But instead of doing that, we were still slapping a visual query interface on top of a database and that's it: the query engine still waits for the entire query to be formulated before it starts processing.

Our first paper about how to change this was a short paper in ICDE 2006[1], long before people were thinking about what we now call *exploiting think time*. That's become a buzzword nowadays, for interactive exploration, but back then, it was hard to get the paper accepted. The reviewers were saying that this is HCI related, so hey, don't submit it to database conferences! As if the usability of data systems were not important! But that's how we database researchers used to view our own field: we were working under the hood, and user interfaces, how the users search or query, is not our problem. That really limited the usability of database systems. I'm glad that that attitude has changed dramatically in the past decade.

> ***If you pick up any […] textbook, its authors will use diagrams to explain the concept of graphs to you. […] But when we are using a query language to search a particular graph data source, we are asking the query writers to write their query in text form, even though that is not intuitive for humans.***

After getting started with human data interaction research through XML we moved on to graphs, where we worked on blending visual graph query formulation with processing. And then we moved on to the idea of creating visual query interfaces in a data driven manner. That means that the user gives us a dataset and then we give them a visual interface for that data, automatically generating it for them. We did this for graph data.

The reason we focus a lot on graphs is because graphs are very intuitive to draw. If you pick up any data structures textbook or any algorithms textbook, its authors will use diagrams to explain the concept of graphs to you. They don't want to describe the graph in text form. But when we are using a query language to search a particular graph data source, we are asking the query writers to write their query in text form, even though that is not intuitive for humans. In fact, humans using drawing to express themselves predates any form of textual expressions.

Suppose that the user has a graph data source and I give them two hypothetical options. Option 1 is that I

[1] Sourav S. Bhowmick, Sandeep Prakash: Every Click You Make, I Will Be Fetching It: Efficient XML Query Processing in RDMS Using GUI-driven Prefetching. ICDE 2006: 152.

give them a box with hundreds of knobs to turn, and I tell the user, "Hey, this is the fastest data processing system you can ever have. Your every query can be answered in constant time. But you figure out how to write your query by tuning these knobs." Option 2 is that I give the user a nice visual interface, and I tell the user, "Hey, I'm not always going to give you every answer very fast, but you don't need to learn query languages. You can just draw your query intuitively, the way you see it, and search." Which one do you think a domain expert will prefer? That's why I believe there's a lot of potential for graph queries to be visual in nature: it's much more intuitive to visualize a structure than to write it down in text form.

What's the state of the art for visual queries today?

Now we know how to process a set of subgraph search queries by interleaving query formulation and query processing. We have a series of papers in SIGMOD[2] that describe how to do this.

We also know how to create the visual interface for a graph data source automatically for large networks, as well as for a large collection of small or medium size data graphs, like chemical compounds[3].

What are you doing with the Initiative on Diversity and Inclusion in Database Conference Venues?

Recently, I was invited to co-lead its subgroup that is looking at issues in managing conflicts of interest in database conferences.

What's a conflict of interest?

A *conflict of interest* is a set of circumstances that creates the risk that your professional judgment or actions regarding a primary interest will be unduly affected by a secondary interest. That's the definition, independent of any specific area. In our context, the primary interest might be an impartial review of a paper, and you can imagine secondary interests as possible benefits like financial gain, favors for your friends, professional advancement and so on.

The secondary interest becomes objectionable when it is believed to be so large that it may have too much influence on your judgment about the primary interest. Essentially, this is nothing but human bias. Conflict of interest is just one aspect of human bias, and there has been a lot of research from the psychology community showing that conflicts of interest indeed have a serious impact on impartial evaluation of someone's work.

ACM has clearly stated a policy about the kinds of conflicts of interest – COIs – to be avoided for ACM publications[4]. ACM also says that knowingly hiding or falsifying the declaration of COIs is actually a violation of the ACM code of ethics.

For a long time, most database venues have had a COI policy written on their websites. When you submit your paper, you should mark this blah-blah-blah and so on, right?

Yes.

These policies are trying to transform the definition of a conflict of interest into a more clear-cut and actionable definition. Now, the problem lies here: how do you check that authors have indeed reported all their COIs?

Why do you need to check that?

For a long period of time, we didn't have to check it because we were operating based on gentlemen's agreements and trust. But I'm not aware of any human society which is purely based on trust and gentlemen's agreements. Otherwise, we would not need any police!

Sure.

The majority of people are law abiding, and the police are designed for the small minority who are not law abiding.

For a long period of time, our community's handling of conflicts of interest was based on trust. We had our policies, and we believed that the authors will

[2] GBLENDER: Towards Blending Visual Query Formulation and Query Processing in Graph Databases. Changjin Jiu, Sourav S Bhowmick, Xiaokui Xiao, James Cheng, Byron Choi. In SIGMOD, 2010.

QUBLE: Blending Visual Subgraph Query Formulation with Query Processing on Large Networks. Ho Hoang Hung, Sourav S Bhowmick, Ba Quan Truong, Byron Choi, Shuigeng Zhou. In SIGMOD, 2013.

BOOMER: Blending Visual Formulation and Processing of p-Homomorphic Queries on Large Networks. Yinglong Song, Huey Eng Chua, Sourav S Bhowmick, Byron Choi, Shuigeng Zhou. In SIGMOD, 2018.

[3] AURORA: Data-driven Construction of Visual Graph Query Interfaces for Graph Databases. Sourav S Bhowmick, Kai Huang, Huey-Eng Chua, Zifeng Yuan, Byron Choi, Shuigeng Zhou. In SIGMOD, 2020.

CATAPULT: Data-driven Selection of Canned Patterns for Efficient Visual Graph Query Formulation. Huang Kai, Huey Eng Chua, Sourav S Bhowmick, Byron Choi, Shuigeng Zhou. In SIGMOD, 2019.

[4] https://www.acm.org/publications/policies/conflict-of-interest

truthfully and completely declare their COIs. Recently, I wrote a piece of data-driven software called CLOSET[5] that is designed to check if this is really the case. We ran it in a few conferences, initially. In a very recent conference, ICDE 2021, we ran CLOSET post-facto and found a non-negligible number of unreported COIs. So far, in every venue where we have run CLOSET, there is non-empty set of unreported COIs.

> ***You need to try to avoid ethical compromises at all costs, because once you're in, you are in forever.***

What kind of conflicts of interest are you talking about?

Those due to coauthorship or being at the same institution.

In the old days, the computer science research community had small numbers of submissions, small program committees, and they could manually check for COIs: "Hey, this guy should not review that guy's paper" and so on.

But we are in a time where in every data-oriented field, PC sizes are becoming huge. The number of submissions is growing like crazy. It is impossible for a PC chair to manually check whether authors and reviewers have declared their COIs. That's totally out of the realm of possibilities. So, we have our conflict of interest policies, and how do we enforce these policies? We have a huge gap. The implementation can't be manual or based on mutual trust anymore because *the data is showing otherwise*.

So, we need tools. We need data driven tools which will automatically check COIs for our program committee chairs so that they can minimize the bias in the reviews. This will have a downstream effect on the quality of the reviews; on the science which comes out, based on which other people build their science; and essentially, on the healthiness of our scientific community. So it's a very fundamental building block that needs to be put in place.

Now, you'll be amazed to hear this next part. We are the data people, right? And we have a lot of data driven techniques.

Yeah.

On a whole bunch of use cases.

Sure.

If you search for the keyword *data driven* in DBLP, you'll get more than 10,000 hits. We have written more than 10,000 papers on a wide array of issues, but very few on the very scientific framework on which we build our science: *our peer review system*. We data researchers really need to work on this because it has significant downstream impact on many aspects of science.

This is where I got involved, because the data was showing that indeed, there are unreported COIs, and we need to have proper management of this, and proper education. Maybe people do not realize the importance of correctly reporting COIs. And we need to look into possible alternatives for defining COI policies, because our current COI policies are not really capturing existing biases effectively.

What have you been doing to fix those problems?

The CLOSET software that I wrote takes as input the list of submissions and reviewer assignments from any conference management tool, like CMT or EasyChair, and determines whether there are unreported COIs. It detects whether there are program committee members whose review assignments violate the COI policy of a venue, and it does this automatically by analyzing data from several different sources. Then it produces a report for the program committee chairs, saying, "Hey, take a look into these papers. Their review assignments may have COI violations, based on the COI policies of the conference."

The major conference submission review management systems like CMT and EasyChair do have some automated conflict detection functionality now. But I hear from the PC chairs that CLOSET is detecting COIs that have bypassed the conference management system's conflict detection system. We don't know how those algorithms work because their code is not published, but CLOSET is regularly detecting COIs they miss.

While we want to detect undeclared COIs, we also do not want to change the traditional way that authors and reviewers and PC chairs interact with the review management system. We don't want to arm-twist them and ask for a lot of data from them. So, one good thing about CLOSET is that it does not request any additional data from authors and reviewers. It just uses the same kinds of data that authors and reviewers have always provided.

[5] https://personal.ntu.edu.sg/assourav/research/DARE/closet.html

The AI community has taken a different approach, where they ask authors and reviewers for a lot of data, Google Scholar pages and DBLP addresses and all recent collaborators, which can be quite annoying. If I'm a senior author and I have a bunch of collaborators, you're asking me to remember all of them and enter them into the review management system? This is too hard.

Taking too much input from users also increases the chances for the data to be dirty. A PC chair recently told me that they tried to collect Google Scholar and DBLP IDs from reviewers and there were a lot of mistakes. So, if we trust this dirty input data and use it to find COIs, we are bound to have a lot of false negatives.

At the 30,000 feet, the COI detection problem looks deceptively simple: you look up the author's name in, say, DBLP, you look at the co-authors list on that page and if the reviewer's name is there then there's a conflict, and if it's not there then there's no conflict. But once you look deep into the data, you find a lot of issues, and the problem is not that straightforward.

The problem is challenging because, first, the data is dirty. Authors do not always specify their names in the same way that their names are listed in DBLP. Second, even DBLP data is not completely clean, although it is a great data source for us and the cleanest one around. Sometimes DBLP has publications of people assigned to someone who is not actually the correct author, mainly due to homonym problems. Because names are ambiguous, DBLP also has disambiguation pages, where many papers end up because DBLP is not certain which person wrote them. There are more than 270 different computer science researchers named Wei Wang in DBLP!

Maybe someday ORCID IDs will solve the problem, but right now many people don't have an ORCID or don't use it when they write their papers. Certainly they didn't use it on their old papers. Many people have multiple ORCID IDs, too.

Because there are so many data cleaning issues, the problem is much more challenging than you would think. Of course, if we just use a string-matching technique, even regular expression matching and try to look for a name, we will always find some COIs. And that gives us that false security that "hey, I can detect that COI!" But the challenge, the actual issue, is the false negatives. What are we missing? We see that when we run CLOSET and detect some of the COIs which were overlooked by industrial strength review management systems.

You were talking earlier about visual interfaces to data. I understand that you are a visual artist yourself and have even sold your artwork to benefit humanitarian causes. What kind of art do you do?

The topics I choose to paint are often driven by socio-political causes, and one that is close to my heart is refugees. We have millions of refugees worldwide from conflicts and other causes. A few years ago was the peak period of the Syrian crisis, as well as the Myanmar refugee crisis.

One of the refugee crises, the Myanmar crisis, was very close to Singapore. A couple of other painters and I were thinking that Singapore is like a bubble, it's a place like nowhere else. People are very well taken care of by the government here, and people don't worry about the basic things in life. Since there is nothing important to worry about, people instead focus on small issues, relatively speaking, issues that probably other countries will not bother to talk about. Being a refugee is not something you see in Singapore and it does not gather much attention from local people. To educate people here about the refugee problem and to support refugees across the world, my art buddies and I had a series of exhibitions.

In the creative process in art, you start by thinking, looking deep into a problem and start observing things the same way as you do in science. You start with nothing, a blank canvas. You observe what's going on in the same way that we scientists observe our data, and then try to create something out of it. And that's what we also do in science.

The painting above tells the story of a boy named Omran, who was found all covered in dust but still alive, in a place in Syria where the buildings were all destroyed by bombing. In the background of the painting you see the total destruction of the city. The T-shirt he was wearing that day is different from the one in the painting, which shows a particular color combination that is a symbol of peace. It was used in a peace flag back when Iraq was invaded, so this color combination means *give peace a chance*. The painting

is symbolic in that the boy is trying to tell you to give peace a chance, through his T-shirt.

The painting above shows refugees who are fleeing Myanmar by boat. In the painting, even though they're fleeing, you can see that they are together. They have nothing but they are together and trying to help each other, as you can see from how they are trying to pick up other people from the water. Some governments were using helicopters to drop aid packages on the sea for them to pick up, and they are swimming out from the boat to get the aid packages and then going back to the boat.

Up on the right-hand corner, there's a dove, a message of peace. And on the left edge, there's a lady who's looking at the opposite direction to where all the other people are looking. She represents Aung San Suu Kyi, looking away from the problem.

Thank you for showing us those.

Do you have any words of advice for fledgling or mid-career database researchers?

I don't want our young people to be like Aung San Suu Kyi, looking away from big ethical problems. I recommend that young researchers be ethics centric and ethics first in their approach to research.

But how can you be acting ethically if your advisor is not completely ethical? What should you do?

There will be some people, I hope, in your university who want to support students who are ethically correct. And if that is the case, one approach is to change your supervisor. That is what our students at NTU do if they find that they cannot work with their advisor for various reasons. You need to take that leap of faith and have the courage to remove yourself from any situations like this. You need to try to avoid ethical compromises at all costs, because once you're in, you are in forever. One favor leads to another and then it leads to another and then you have a big problem.

Thank you very much for talking to me today.

It's been a pleasure!