Transcript

[00:00] [music]

Derek Bruff: [00:05] Welcome to "Leading Lines," a podcast from Vanderbilt University. I'm your host, Derek Bruff, director of the Vanderbilt Center for Teaching. In this podcast, we explore creative, intentional and effective uses of technology to enhance student learning, uses that point the way to the future of educational technology in college and university settings.

[00:23] In this episode, we feature two interviews conducted by Cliff Anderson, the director of Scholarly Communications at the Vanderbilt Library. Cliff is a librarian and a leader here in Vanderbilt's digital community's efforts, and he's always bringing to my attention powerful new tools for scholarship and teaching.

[00:37] Both of his interviews focus on one such tool called Neo4j, an open-source platform that can be used to visualize and analyze data and connections among data.

[00:46] Cliff interviews his Vanderbilt Library colleague, Suellen Stringer-Hye, who has an awesome job title, Linked Data and Semantic Web Coordinator. Suellen has worked with a number of faculty members and students here at Vanderbilt, helping them use Neo4j in their research.

[01:00] In the interview, she talks about some of those projects and how a database tool like Neo4j can be easier to use than one might think.

[01:07] In the second interview, Cliff interviews Michael Hunger, who handles developer relations for Neo Technology, the company that developed Neo4j. Michael shares a few more examples of how Neo4j has been used, including, as I learned listening to the interview, in the analysis of the Panama Papers, a set of 11 million financial documents leaked earlier this year.

[01:27] First, however, Cliff talks with Suellen Stringer-Hye here at Vanderbilt about Neo4j and about the kinds of graphs that it uses to make sense of data.

[01:36] [music]

Cliff Anderson: [01:39] Hi, this is Cliff Anderson and I'm here for the Leading Lines Podcast. I'm the director for scholarly communications in the Vanderbilt University Library. I also work across the campus with our friends in supporting education technologies in various fields.

[01:56] I'm here today with Suellen Stringer-Hye, who is our coordinator for Linked Data and the Semantic Web in the Vanderbilt University Library.

[02:04] Good morning, Suellen.

Suellen Stringer-Hye: [02:06] Hi, Cliff.

Cliff: [02:08] I just wanted to ask you some questions about the way in which you've been using graphs, in particular, to teach students and how that fits into our overall education technology program. Before we get there, why don't we just start introducing who you are and how you got to the point where you are in librarianship?

Suellen: [02:32] OK, great. My name is Suellen Stringer-Hye, as you said, and I've been working in the Heard Library now for about 20 years. I started off working in Special Collections and then, when the new online catalog was coming on, that was Acorn way back when, I started getting interested in technology and so I started working with that wing of the library.

[03:03] It was at that time called Library Information Technology Services. Worked for many years doing Web development and support for various of the services that the library provides electronically. When Cliff came on board for the scholarly communications team, it opened up a window for me to do what I've been interested in in a long time, which was work on linked data and the Semantic Web.

Cliff: [03:37] In fact, I think we might mention that the way that I've discovered your interest was reading some articles you had written on a subject called topic maps, which were a really interesting technology that never really took off, but thinking about how to model objects

and relationships using graphs in XML.

Suellen: [03:57] Yeah, it was probably the background that gave me…I had become fascinated by going to some XML conferences and meeting some people there who were developing XTM that stands for XML for Topic Maps. That was developed out of some people who were working on back of the book indexes and they were trying to come up with an XML serialization that would allow them to do the kinds of things you do with back of the book indexes, electronically or digitally.

[04:40] It was really fascinating, it just for various different reasons never really took off. RDF supplanted it in a lot of ways, but it did give me the background to understand how valuable it can be to use graphs for various educational technologies.

Cliff: [05:00] Let's talk a little bit about that. I think one of the interesting things about graphs in particular is that they're really everywhere. In thinking about the use of graphs as an education technology…first, I think we should probably put a caveat, and you can make this point better than I, that we're not talking about bars and charts here. We're talking about something else.

Suellen: [05:28] We're talking about what's essentially a mathematical graph, which means it has a lot of nodes and edges. Basically, and back to the topic map model, you have topics or objects, and then relationships between those.

[05:48] The reason that that is so useful, in terms of modeling information, is that in the past we used to have to model information with embedded relationships because we could use hierarchies, but we never really could code the interconnections between things.

[06:06] I often say it's a lot like in the old days when you'd go to the card catalog and you'd get to see also reference. In order to find and that see also reference, you pick yourself up and walk over to the other place where that was. This is just a way to code those relationships so that we can do various things with them.

Cliff: [06:31] What I think is interesting is once you start thinking along the lines of a graph model, you begin to see that lots and lots of things that we encounter in our daily life can be modeled easily as graphs. I think that's where it comes in for an education technology. This is a technology that runs the gamut.

[06:51] It's used…modeling as graphs, used very heavily in natural sciences, also in the social sciences for things like network analysis. Also, in particular, the digital humanities. I think that might be an interesting place to start because you've been teaching several workshops in another context where you teach people to apply graphs to what are classical humanities problems.

Suellen: [07:19] Yes. We have used graphs to analyze things. An example of one that we did was we took a corpus of letters from Flannery O'Connor that we have at the library. Then we reached out to some other libraries to get some metadata about the letters that they have.

[07:42] Then we were able to encode that and create some other relationships between the dates and the subjects and who she wrote to, and things like that. By creating that particular graph, we were then able to ask questions of it that would have been much harder without that particular way of going about it.

Cliff: [08:07] Tell me a little bit about the tools that you're using to do this kind of encoding as a graph.

Suellen: [08:14] We have been using a tool called Neo4j, which is a NoSQL database, which just means it stores the data in graph format with the nodes and edges. There's several ways that you can use Neo4j. One is as a large scale database. You can upload Excel spreadsheets into it, then do some networks analysis on it.

[08:45] Neo4j has lots of different levels that you can work with. It's really useful, but one of the things that we have found most useful and where I have found most useful in teaching is something called a GraphGist. The GraphGist was created by Neo4j just to get people to be able to try out what it means to encode information in a graph and then be able to query it.

[09:17] It's very simple. It's very easy to work with. I found that humanities scholars, it's just very intuitive to them. They can understand and that a lot of the problems that they're interested in or want to research lend themselves well to graph format. Then on top of that, they don't need a programmer to begin to work with this tool. It's really helpful in a lot of different ways.

Cliff: [09:46] One of the things that's really interesting about Neo4j is the query language, Cypher. What's nice about Cypher is that uses a visual programming style, based on ASCII art

syntax. Basically, you model a node with two parens surrounding the node. A relationship looks like a dash with a caret pointing to the right.

[10:12] One of the things that's nice when you're teaching humanists who may never have programmed before is that the syntax is very intuitive and they seem to grasp it very quickly.

Suellen: [10:24] It's actually fun to work with. I've had a lot of people say they really actually enjoy doing that kind of coding because like you say, it's very intuitive and it's visual and it's just really interesting.

Cliff: [10:40] Can you tell us about some of the other Neo4j products that you've worked on?

Suellen: [10:44] I've done a little bit working with some short stories. Two short stories come to mind. One was a Flannery O'Connor short story called, "The Life You Save May Be Your Own."

[10:58] I also published an article that used a graph to illustrate some points about a Vladimir Nabokov short story called, "The Vane Sisters," and it was the type of thing that lent itself very well to a graph because it operates on several levels.

[11:19] I was able to use the graph to point to some outside references or resources that it was fairly clear that Nabokov used to illuminate the story. The graph allowed me to do things that I wouldn't have been able to do as easily with expository writing.

Cliff: [11:41] I think the Nabokov example is particularly interesting because as I understand from you, his stories, at least some of them, are not self-contained. They actually make references to things in the external world that are very useful to shed light on what he's talking about.

[11:58] Being able to model his stories as a graph, and be able to point beyond the story and then reach back into it is actually an important form of literary analysis for Nabokov.

Suellen: [12:07] Very, very true. Yes.

Cliff: [12:10] What are some of the hardest things for humanity students coming to grasp for

the first time? Where do you seem them having the hardest time?

Suellen: [12:17] I think the hardest time for them is getting over the idea that they can't code or that they're not coders or they're not programmers or that's not just their thing. Once they get past that, they seem to be able to very easily do the data modeling, which is necessary because that's just the way they're used to thinking about things.

[12:43] It's a lot like those brainstorming bubbles when you're trying to decide what you want to write a paper about, or whatever. You just draw a bunch of circles and then connections between those and when you get done with it you go, "Aha! I know what I wanna write."

[13:00] The process is very similar when you're doing data modeling for use with graph analysis. That part is not a problem. The only real problem is the idea that, "I don't know how to do this, this huge, steep learning curve." I think once they get over the fear of it, they really take to it pretty easily.

Cliff: [13:24] We should also mention that you regularly teach workshops. You'll be teaching a workshop in the fall on Neo4j and probably, likely at our THATCamp. You've also recently published an article on the library website about getting started with Neo4j.

Suellen: [13:39] Yes, I put up a little tutorial. It's under our workshops and I think it's called, "Getting Started with GraphGists." I walk you through the steps that you need to take in order to get one going.

[13:55] That one we used ancient deities or Greek gods, ancient gods. You can see there's a lot of different applications for using graphs in the humanities from literature to history to all different applications.

Cliff: [14:21] I think there's room, the article that you mentioned about Nabokov, I believe that you wrote that as a GraphGist. I think that there's room for humanities students to begin exploring writing articles in his style where they mix graph code along with their narrative, expository text.

Suellen: [14:40] There's a lot of different applications for it, you can publish, and then another useful thing is just to collect your research because I think we all know, any of us who's done humanities research, end up with a lot of post-it sticky notes throughout your

books and graph pads full of notes.

[15:02] You're always paging back through to try to figure out, "Where did this come from?" If you take the time to encode that into a graph, then you could use it to help structure your research. It could be useful just as a research tool, as well.

Cliff: [15:24] This is terrific Suellen. I really appreciate you taking the time to talk about this. We're asking all of our guest to…since we're all talking about digital experts, we thought it'd be interesting to ask what is your favorite analog education technology?

Suellen: [15:42] Being a librarian, of course I love books. I'm the first person to really like to get off the computer and be able to touch the pages, going to walk away from anything that's plugged in or connected and go sit outside, listen to the birds and read that way.

Cliff: [16:11] Suellen, thank you so much for this conversation. I think it's really interesting to hear what you're doing with graphs and how you're teaching students to think about their subjects with new digital spectacles on.

Suellen: [16:26] Thanks, Cliff, I've really enjoyed talking about them. As you can see, I'm really…maybe I wouldn't go so far to say passionate, but I really do like working with graphs. They're really fun, interesting, lots of different uses, and I like getting people excited about them.

Cliff: [16:45] OK, we'll talk again soon, I hope. Thanks a lot, bye.

Suellen: [16:48] Bye.

[16:49] [music]

[16:52] Hi Michael, this is Cliff Anderson calling, and we're really glad to have you as guest on our podcast. We're really excited to talk about Neo4j and its uses in educational technology. Maybe we could start by having you introduce yourself and just tell us a little bit about your role at the company?

Michael Hunger: [17:12] Yeah, Cliff, happy to be on this show. I'm really excited to talk to you. I've been with Neo4j and Neo Technology, the company, for a while. I joined in 2010

when we were still 10 people and now it's been growing to almost 120.

[17:30] In all this time I've had a lot of different roles, but mostly I'm taking care of people using Neo4j. I'm calling myself caretaker general of the Neo4j community, which describes nicely that I'm using my time to help people be happy and successful with Neo4j and doing it with a lot of different means, from articles and blog posts and books to conference talks and meetup events, one-to-ones, hands-on support on the convention floor, and other places.

[18:04] Also giving people ideas. I'm really good at creating tons of ideas and then just spreading them out and exciting people for what's possible with this kind of technology. In general, I'm jack of all trades at a company, doing a lot of connecting of people inside and outside of the company.

[18:23] I think that's very important. If you have such a big field as information technology, that you bring the right people together, because I think between interested parties that really great projects emerge.

Cliff: [18:41] Maybe you could give us your elevator pitch for what is Neo4j, for those who haven't heard about it yet.

Michael: [18:48] Yeah, I'll do that. Neo4j is a database. Like other databases as well, it's meant to manage and store and process data, and make it possible to create this data. But unlike other databases, it focuses on connected information.

[19:05] As you probably know, there is no such thing as disconnected information in the real world. Everything is connected with everything else, more or less. Many other databases are not really good at handling this connectedness of data.

[19:18] Neo4j was built from the ground to do this in a really efficient manner, to store, handle and make the data available quickly and efficiently, as soon as it has a lot of connections.

[19:32] If you had experienced a lot of problems in other databases, for instance, to get really differently structured data into the database and then quickly out again, especially if you have queries or questions that involves joining together many, many tables, then Neo4j could be a tool for you.

[19:50] In general, also the model that Neo4j employs, the graph model, which is connected entities, so entities connected by relationships, is such a suitable and versatile model that you can use it for almost any kind of domain.

[20:07] From things like cancer research, to managing curriculums, to recommendation engines, to routing, to even something like logistics and package management, and many, many others as well. It's actually not limited by the technology, just rather by your imagination and the problems that you want to solve.

Cliff: [20:33] One of the things I love about the way that Neo4j talks about data modeling is that it can capture your whiteboard sketches, so that when you're thinking about a domain and you start putting up circles and drawing arrows, more or less you can take that model and encode it directly in Neo4j.

Michael: [20:49] Yeah, that's true. This whiteboard friendliness is a really big advantage of the technology because with other databases you have the problem that, at a certain point, only developers can work with the database and handle the data.

[21:04] But what we try to do is to keep everyone who's involved in the project in the loop all the time, from people that come from the business side to people that are users. Because this whiteboard model is so natural to humans, we start to doodle on whiteboards or blackboards as soon as we discuss any kind of topic. It's really nice to see that people can always get along.

[21:27] Even if the data is stored in a transactional database, like Neo4j, they still see the circles and arrows all the time and can reason about it and can spot issues or find new connections and so on.

[21:43] The other thing that we also try to do is not just do it on the whiteboard level and on the visualization level, but also with our query language, Cypher. We try to take this whiteboard modeland the patterns that you draw on whiteboard, and turn them into a textual queer language.

[22:00] What we did there is actually to use ASCII art to represent these two-dimensional drawings in a one-dimensional text that works actually quite pretty well. What we actually do is we surround entities, or nodes, how we call them, in round parenthesis, so they form a

circle, visually.

[22:20] Then use something like dash-dash-greater than or dash-dash-less than to indicate the arrows. You can, easily, if you look at a query in Cypher, still see the original patterns that you were looking at or talking about.

[22:37] For instance, if you're asking for an author who also the book which had characters appearing in the book, which enveloped certain things, for instance, you have this kind of pattern that you would usually draw on the whiteboard.

[22:51] You more or less draw it in text. Because that's so visual and also efficient, some people even say it's more expressive than English in declaring these connected patterns of information, that you then let the database figure out how to get them quickly back to you, and aggregate and project information that you collect while searching for these patterns in the data.

Cliff: [23:20] You described Cypher, which I think is one of the real joys of Neo4j. I think you've done a wonderful job of developing that language. We've had the experience here at Vanderbilt of using Cypher in workshops, and in the classroom, and on student projects, and one of the things we find is that, as you say, students really take to it because it does have a visual component that models the types of graphs they're building.

[23:49] The other thing that you've done that I think has been terrific for a teaching tool is this concept of a GraphGist. Do you want to say a little bit about what GraphGists are?

Michael: [23:59] Yeah, definitely I'd like to. In general, I think Neo4j, or the people at Neo4j, from the very beginning spent a lot of time and effort on documentation, making it as easy as possible for people to get started with Neo4j, but also to have documentation about common graph models or typical patterns that you use the graph for.

[24:23] When you are looking into that to bring it to a broader audience, it actually turned out that with Cypher, they have the ability to take regular text files, so you have headers, you have prose, you have images, you have tables.

[24:41] You can intersperse that with code sections that either create data or query data. Then we can take one of these text files, which is just a regular text file that you can store on your

computer, on Dropbox, in a version control like GitHub and so on.

[24:58] Then to take this text file, extract the statements, run them against a temporary Neo4j database and then render the graph visualizations or the table results that come back from these queries. You have a live document that represents your ideas that you've described in prose or formulas or other visuals.

[25:23] Within that document you have exactly the same queries that create the graph data and create the graph, too, so that you have a live document of what you want to do and then, also, demonstrating how you want to do that.

[25:35] Over the years, we had several challenges for people to produce these GraphGists, and also a lot of people actually creating them on their own. In total, we have now a collection of, I think, almost 300 of them across a lot of different categories, from things like humanities, to healthcare, to logistics, to recommendations, to storytelling.

[26:01] If you want to see some of them, you can go to Neo4j.com/GraphGist. You can find a rich collection of these really interesting documents. As an author, you can also run this locally, so if you want to keep your data inside your company, then you can also say, "OK, I'll just run this locally."

[26:23] The use for students is actually quite interesting because I'm also doing classroom lectures sometimes, either directly or remotely, and I even use GraphGist as homework assignments for students.

[26:36] In the lecture they learn about graphs and Neo4j and Cypher, and then they give them GraphGist template that they can then put in a model that they like, either something that they are interested in, a hobby or a project they've been working on, and create a little bit of data, write one or two queries, and then hand this in as a homework assignment to the professor.

[27:01] So far, the professors and the students that we did it with actually liked the idea and we got some really interesting content out of that because if people get started once and they really care about the subject, then they go to very great lengths and depths to create something that's truly, truly amazing.

[27:21] Plus, it's always astounding, the variety of things that people use Neo4j and the graph model for, so it's a really exciting time to use. Every day you learn something new, how someone used this technology. That makes it really a lot of fun to work with the people in the Neo4j community as well.

Cliff: [27:43] We'll put the link to the GraphGists in our show notes, but it is really a wonderful collection, especially if you're thinking about a domain that you might like to model to see how others have done it. I love the idea of using it for homework. We'll have to try that here.

[27:58] One of the other things that you've done that I think has been really great for helping people to get started is that you've built in a visualization engine into Neo4j, so that it's not just coming out as tabular data but it also got a graphical representation from your queries.

Michael: [28:17] Yeah, that's true. I think that ties to what I said before, that you want to keep this rich, whiteboard visualization of your domain. Not just on the whiteboard and in the query language, but also in the output that you get back from your database.

[28:31] As in the database, the data is stored as nodes and relationships anyway. Putting them on the screen again is just a little step but very, very useful because many people have to take this visual cue and have a much easier time to work with the data if they can see it.

[28:51] Even if you can't visualize the whole graph, because many graphs go into the billions of nodes if you have real-world production data, but you can always visualize a sample or a small part of that and explain something.

[29:08] Also navigate visually, you can double-click on nodes and then they expand with their neighborhood and so on. I think this aspect is really important and many other tools in the library that people also use to integrate these kinds of visualizations into their own projects, which makes it really easy for them to also have, within their own application, a graph visualization that allows them to bring home the richness of the domain, which is, if you put it in table, it's just lost.

[29:43] Because in tables you just have columns and numbers and you actually don't see the forest for the trees anymore. I think the graph visualization is really an important aspect, as you said, to make the data accessible again for people.

Cliff: [29:58] If you were to come here to Vanderbilt, you'd see that we've actually put several on display in our exhibits. The students who have worked on these have been very proud of the visualizations they've produced.

[30:10] Again, I think one of the things is it's open, you can add something like D3.js to do more complicated visualizations, but you've made it very easy for people without a programming background to get started. Then when they need to scale up and do something different, they can do that.

Michael: [30:27] Yeah, that's true. You've probably heard of the Panama Papers investigation this year? The ICIJ, the Investigative Journalists Consortium, also use Neo4j as the underlying technology to connect all the different bits and pieces of information.

[30:44] For them, they had quoted, "It's like magic." Put it in there and suddenly it's available as a connected network of information that you can see and you can see connections between people and events and companies and so on.

[31:00] As soon as you start to add new connections then new patterns emerge and you have for instance, two or three of the four hundred journalists that look on the data. Add new connections and suddenly a new story emerges for a third journalist.

[31:15] Having this available for everyone is really powerful because then you cannot just take the typical business applications, but you can also do so much more. There are projects like the-codex.net (http://the-codex.net/), which uses renaissance letters from da Vinci and Michelangelo from Italy to draw a full picture of timeline and places and events out of the renaissance.

[31:47] Or if you take other data sources like information about authors, like the work of an author, or if you take full libraries, for instance, and cross reference to the works in the library, then I think all this information is so much more accessible than it is in an index or in a tabular database, I think.

Cliff: [32:12] I think it's wonderful that you've also made that available as a docker image so that people who want to explore that data easily can get started really quickly. I was at a meet-up in San Francisco where they were helping people hands on to get that docker image up and running.

[32:33] You've done a great job packaging this so that people can easily get started with queries and exploring without a lot of preliminary set-up. Again, thanks Michael for talking with us, and as we ask all our guests, what would you consider your favorite analog technology?

Michael: [32:51] That's an interesting question. I'm a big fan of da Vinci. Actually, he has a lot of really good exhibition just a few weeks back and I think, looking at all the things he invented and built, just the simple mechanics, I think that's my favorite technology.

[33:10] From his flying machines to all the levers and pulls, all the things that made work so much easier for people, I think that's the thing that I really like most, to see what's possible if you're just smart about things and not just use brute force.

Cliff: [33:34] That's a wonderful answer. I love those notebooks too. We really appreciate the time you've taken with us and we really appreciate everything that you've done at Neo4j to help students get started with thinking about graphs and databases. We look forward to seeing exciting things to come. Thanks so much again.

Michael: [33:55] For everyone who is able to, I invite everyone to come to our conference in San Francisco to GraphConnect. It's on October 13th.

Cliff: [34:05] Awesome. Thank you so much. Thanks again Michael, I really appreciate it.

Michael: [34:09] Thank you.

Cliff: [34:10] Bye-bye.

[34:10] [music]

Derek: [34:10] That was Michael Hunger of Neo Technology, and earlier, Suellen Stringer-Hye, a librarian here at Vanderbilt University. You can read more about both of our guests by following the link in the show notes, which include a link to the "Getting started with GraphGists" tutorial that Suellen mentioned.

[34:27] Thanks to Cliff Anderson, director of scholarly communications at the Vanderbilt University library, for those interviews.

[34:33] You've been listening to Leading Lines, a podcast on educational technology from Vanderbilt University. That podcast is produced by the Center For Teaching, The Vanderbilt Institute for Digital Learning, the Office at Scholarly Communications, and the Associate Provost For Digital Learning.

[34:46] You can find past episodes on our website, leadinglinespod.com, and you can follow us on Twitter where our handle is @leadinglinespod. Look for new episodes the first and third Monday of each month.

[34:58] I'm your host, Derek Bruff. Thanks for listening.

[35:00] [music]