SDS PODCAST EPISODE 923: GRAPH ALGORITHMS, GRAPHRAG AND CAUSAL GRAPHS, WITH GRAPH GURU AMY HODLER

Show Notes: http://www.superdatascience.com/923

Jon Krohn:	00:00:00	Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today we've got an episode for you all about graphs, no, not plots, but the powerful graph data structure with many time graph analytics, book author and in general the Worldclass graph analytics guru Amy Hodler. In today's episode, Amy introduces the graph data structure, graph algorithms and all kind of cool new graph applications including graph rag graphs for LLM memory and causal graphs. Amy is absolutely terrific at explaining these concepts in an easy to understand way, so enjoy this one.
	00:00:35	This episode is Super Data Science is made possible by Dell, Intel and the Open Data Science Conference.
	00:00:40	Amy, welcome to the SuperDataScience podcast. Where are you calling in from today?
Amy Hodler:	00:00:47	Well, thank you for having me. I am calling in from Little Kettle Falls, Washington.
Jon Krohn:	00:00:53	Little Kettle Falls. What's it like at Little Kettle Falls and is there a big one nearby?
Amy Hodler:	00:00:58	There is not. There are under 1700 people in this small rural area of northeast Washington, so a lovely, beautiful area, but definitely small town.
Jon Krohn:	00:01:10	Nice. Lots of outdoors I guess.
Amy Hodler:	00:01:11	Absolutely. We hike every day.
Jon Krohn:	00:01:13	Lovely. That sounds nice. Living in Manhattan. It is something I literally yearn for. I sometimes lay in bed just in the middle of the day and imagine that I'm under a canopy of leaves. There's something soothing about that.

Amy Hodler: 00:01:30 Oh, well come on out. We we'll do some kayaking, some biking and hiking.

Jon Krohn: 00:01:35 Oh my goodness. That sounds so great. Now the real purpose of you being on the show is not to discuss the great outdoors, as wonderful as it is, but to talk about graphs, which you are such a deep expert in. And so first off, I guess we should kind of explain what graphs are quickly off the bat just to make sure people aren't thinking about plots.

Amy Hodler: 00:01:55 Yeah, absolutely. That's usually one of the first questions are what are graphs? So graph is actually a term that comes from mathematical history, going back to the 17 hundreds actually. So graphs have been around for quite a long time and it's basically a way to capture relationships in data. And so if you think about when you go to a whiteboard and somebody asks you to draw out your process or your organization and you do circles for the nouns and you do lines between them, and those are usually the verbs, those are the relationships. That is all a graph is, it's just a way to capture entities, think of them as nouns and relationships. Think of those as verbs. So that's simply what a graph is.

Jon Krohn: 00:02:43 Very nice. And for our data scientists out there or as statisticians out there, the dots in the graph are nodes and then the connections are edges. Are there any other key terms that we need to know in world?

Amy Hodler: 00:03:00 I would say nodes are also called vertes or vertex if you will. And your relationships can be edges or links. So there's a lot of different terminology, but the feeling is all the same. Sometimes people do talk about property graphs, which are for your data scientists out there are graphs that have properties on them. So you can have your nodes and your relationships can have multiple properties and it has a very hierarchal feel to it, at least to

me. And then people sometimes will talk about RDF or resource description framework or triples, and those are a little more verbose way of talking about nouns, verbs, and objects. So subject object predicate is what you'll usually hear about. Those are just different ways to model those relationships. So those are the other terms that you hear most network science, of course.

Jon Krohn: 00:03:56 That's cool. Tell me about those things. Tell me about RDF in a bit more detail. You talked about object predicate. Maybe you could give me an example so I can wrap my head around that.

Amy Hodler: 00:04:07 So it's just a different way to think about, it's another way to generate a sentence, if you will, or a fact. And so you might have something where Amy works at company and you can basically create that with a triple, so you have Amy and then the concept of works at, and then you have the company or whatever the company name is. And so that's the original way that a lot of knowledge graphs came about. It has to do with the semantic web. So it's been around for a long time and there's a lot of academic research around that tends to be more verbose than your property graph. And so depending on different people needs. So depending on your needs, you might go with something that is based on a resource description framework or RDF or you might go with something that's a property graph which has a little more flexibility. I would say with RDF, there's a lot more enforcement of rules and logic, and with a property graph there's a lot more flexibility. So depending on the team's need, you might go with something that's highly flexible but has less enforcement of logic and rules, which can be good, but it can also get your team in trouble or maybe you need that sort of flexibility. So it really just depends on team needs.

Jon Krohn: 00:05:35 And so something in the little bit of graph theory that I know, is it right to say that one of those properties can be

like a quantity? So for example, you could imagine your social media network where everybody that you are connected to on LinkedIn, you're connected to them by an edge and everyone including yourself is a node. You have these edges connecting you. And then you could say you could have a quantity, I don't know if this is the same as a property, but you could have a quantity on each of those edges, which is how many times you've viewed their page or how many times you've commented on a post of theirs or something like that.

Amy Hodler: 00:06:16 So the properties can be just about anything. So at least in a property graphs, and that's part of that flexibility. So it could be anything from a descriptor like a color of a car to the make and model to the year to a quantity. It can be a strength. So it can be, and in fact, that's one of the ways people get around some of the explosion of data is to just do a aggregation of the number of times you've looked at something. So you can basically create a strength. It can be zero to one if you need to normalize something and you need to develop a strength on, especially on the relationships, a lot of times you'll see that strength, but it can also be a geospatial, so you can put a geospatial code in there. It can be a lat long if you are doing geospatial, it can be a string, a number, various different data types as well. And even sometimes a list depending on what graph you're using, so you can use it to then refer to other things. So yeah, lots of flexibility in what the properties can be.

Jon Krohn: 00:07:22 How did you get so into graphs, Amy? I mean, you're the founder and executive director of a company called GraphGeeks, which specializes in bringing together people that relish connected data. So I guess people who love network science, as you referred to it earlier in the episode, how did you get so into?

Amy Hodler: 00:07:40 Graphs? That's an interesting story. So by accident, what's the story? The way you fall in love with anything is while you're doing something else. So I was actually working at Hitachi with the IOT group and what we were noticing was that the end devices had unusual behavior. So I would say I thought it was emergent behavior, it wasn't, but it was this behavior of end devices when they started to interact with each other. That led me to complexity studies and network science because that studies how things interact together, whether you're talking about brain banks, IOT devices or cars on a freeway, you can network look at any of those as network. And so starting to try to understand why these edge devices were having unusual behavior when they worked together led me to start looking at network and network science. And then how you deal with a lot of network science concepts is using a graph because until you can represent your network as a graph, you then can't use.

00:08:55 That's when you get to use computer science capabilities because you have to model it in a way that computers can deal with it. So that led me to taking classes in graphs just to understand them. I had a good friend that was working atray at the time, the old supercomputer company, Craig, and they had a graph engine and he picked up the phone one day and said, Hey, I think you might be interested in this graph stuff. What's that about? Tell me about it. And I just never looked back because, and I will say once anybody who's interested in graphs and network science, once you start seeing the world as a complex relationships between things, you can't unsee it. So obviously the foundation of most of what the world is.

Jon Krohn: 00:09:48 So when you're hiking around in Washington state, you're supposed to be relaxing, but you're really just kind of making graphs in your mind.

Amy Hodler: 00:09:56 And you're saying that's not relaxing. Yes, I think about graphs and networks all the time because how do you not think about relationships between things?

Jon Krohn: 00:10:08 Cool. Alright, so let's talk about some real world use cases. You actually, before we started recording, you were telling me about how with crime that is often a really good use case of graphs. It can give you insights that you couldn't get with any other kind of approach.

Amy Hodler: 00:10:25 And I would say finding bad behavior, whether we're talking about money laundering, that's a really typical one. Fraud, fraud rings, but also very international criminal organizations as well. Even things like supply chain crime as well, which is actually quite significant. The reason why graphs are so good about finding aberrant or bad behavior is that if you think about how somebody trying to hide their tracks, if you have activity going on, usually there's multiple touch touchpoints in criminal behavior. So it's not like somebody walks through the door and advertises that they are doing something criminally bad, they're actually trying to act and behave in a way that is normal to other customers or patients or what have you. And what you really need to do is be able to understand, connect the dots between multiple behavior and touchpoints over time, which is if you imagine what that would look like, that looks like breadcrumbs.

00:11:35 And you can almost, I can see the graph in my head, you have these dots of touchpoints and links between them relationships and you breadcrumb through that to see aberrant behavior or an aberrant or an anti-pattern or pattern that shouldn't be there. And then the other thing is those things usually happen over time, so you can't just snapshot in and look for a bad actor. You have to be able to look at that over time and look at how the behavior develops over time. And the other thing is really

bad behavior is rarely a one person or one entity point in time thing. It usually has multiple different collaborators. And so again, you need to be able to see the relationships between those collaborators and their behavior and their addresses. And so you can imagine that gets into lots of dots and links between them and that's hard to look at as rows and tables. If you look at that as in rows and tables, you're just not going to see the pattern and graphs have that unique ability to pull that important pattern out so that we can see that

Jon Krohn: 00:12:42 You explained that really well, that was really easy to understand and gripping frankly. What other kinds of use cases are out there in addition to crime?

Amy Hodler: 00:12:53 Well, so recommendations is a classic one as well. So for example for Netflix, if they're recommending a movie to me that I haven't seen, but it's a movie that has a director in it that I like or a director in it that I maybe like, but the other people who like that director recommend this other movie. And so you can build a similarity based on behavior and what people like and don't like. So that's classic recommendations. You see that all over the place. The other thing you can do things like optimizing networks of things like supply chain. We all saw supply chains break down during COVID totally broke down. And part of that was because most of the way we were looking at supply chains was looking at the past, and this is the classic machine learning mistake is that you do correlations, you make predictions based on what you've seen in the past, but if you don't understand your network, your customers, your supply chains, your partners very well, and you're just predicting based on what you've seen in the past, when there is an event that shifts the underlying reality, you are going to fall down, you're going to make wrong predictions.

00:14:05 And so you need to be able to understand the patterns and understand how things shifting in the supply chain may shift your predictions. And you can do things like finding the optimal route. So graphs historically, and I think World War ii, they started being used for supply chain looking at supply chain during actually troop supply chains during World War ii. And if you have something that blows up or you have a railroad that can no longer be used, you need to look at another route. And so graphs are really common for route planning. Uber uses graphs, things like that. How do you find the best route? It's all through different points, but in a supply chain standpoint, if you a warehouse that goes out of commission or a port that you can no longer dock in, what's your next best route? And that is a common graph use in supply chains is the top K or the top routes that are optimized for these complex shipping and delivery routes. So that's another really common one would be supply chain. But quite frankly, anytime you have connected data you can use graphs to help you optimize.

Jon Krohn: 00:15:20 I'm starting to see that how there are graphs everywhere and how you are seeing them out on your hikes, crime and supply chain, all kinds of real world use cases you hear sometimes I hear maybe you don't hear, sometimes I hear complaints that things like graph traversal, so doing operations over graphs are tricky, can be computationally expensive maybe relative to some other kinds of approaches. What do you say to that? And maybe this is also the time to bring up that you wrote a great or co-authored a great O'Reilly book called Graph Algorithms. Is that about kind of graph traversal and making the most out of this data structure or is it something

Amy Hodler: 00:16:01 Else that book the graph algorithms book that I co-authored for O'Reilly and graph algorithms in general, it's focused on, I would say the overall algorithms that are

usually a little more holistic graph analysis. So trying to understand your graph and pull out important information. Graph traversal has two elements. You do see graph reversal in graph algorithms, but you also see graph traversal in ad hoc queries. And I will say in ad hoc queries they can be tricky because you don't know how somebody's going to ask the question. And that can be very tricky because if you imagine a pairwise question, the classic joke is amongst graph people, one of the classic jokes is you wouldn't believe, again, I had somebody ask if we could calculate the shortest path between all pairs, all pairs. Shortest path is the classic algorithms and the keyword there is all.

00:17:06 And if you imagine looking for the route between every two pairs of nodes in your network and comparing all of them and looking for the shortest that is computationally crazy, you wouldn't do it. There's a ton of tricks to get around that, but I think graphs can be computationally complex and if you approach it with a very naive sense of I'm just going to ask everything I can ask, I'm going to ask for all pure shortest path, you will have problems. So you have to have, I guess just some thought it's very flexible. You can do just about anything and therefore you are empowered to do things you shouldn't do. So it's just coming with a little bit of thoughtfulness on do you really need all pairs? No, you probably don't. You probably are looking for all pairs among the top K. That's an easy way to get over some of that computational complexity. The other thing is with the ad hoc queries, if you can figure out what are most common queries that are needed with your organization, you can then tailor those and optimize those to both your real needs, but also you can tailor your data model. So your data model can impact your query timeframes as well. So they're different model, they're a very powerful model, but it does make sense to have a little more thoughtfulness when you are approaching whatever the business problem might be.

Jon Krohn: 00:18:38 Cool. So I got an insight there perhaps into another way that data can be stored in a graph format just by you saying the question shortest distance, it implies that there's not an arbitrary distance or that they're kind of equidistant. It sounds like you could kind of have a location in a two dimensional or three dimensional or many dimensional space for each of the nodes. Is that right?

Amy Hodler: 00:19:05 Yeah, you can look at graphs as a dimensional space both. It's funny when I talk about how far away things are, I think of the number of hops

Jon Krohn: 00:19:17 I see.

Amy Hodler: 00:19:19 But there's also a distance. If you imagine a relationship you talked earlier about or you asked earlier about having numerical values on relationships, you can easily do that as well. And so you might actually have a numerical distance on your relationship that is actually related to a physical distance or often it can be related to a time cost in the supply chain. How much does it cost to go between this port and that port? Those can all be considered distances. But yeah, when I'm thinking about computational overhead, I'm thinking about number of hops and hops. So you go, you hop from one node to another, from one person to another. Like Jon, you and I were introduced, reintroduced by somebody else, we would've been a hop out. So we weren't a direct connection, but we had a hop to go through. And so you have to hop between those and that's where the computational complexity can come in. Those hops are often in a relational world, are often joins. And so if you have a lot of hops in a relational world, you have a lot of joins and that can also be computationally complex.

Jon Krohn: 00:20:29 Right? Shout out to our mutual node. Absolutely. Michelle Yee, who was an episode nine 15, absolutely

extraordinary individual. She was. It's a great episode for people to listen to. Yes, absolutely. Yeah, she's like you outstanding at explaining concepts and funny and warm. Both of you share a lot of those attributes. We could list those as, what is it again? When a node can have

Amy Hodler: 00:20:59 Properties,

Jon Krohn: 00:20:59 Properties.

Amy Hodler: 00:21:00 Properties. Okay. Funny story about Michelle. Michelle and I love to present together, especially in person. We started off when we got to know each other, I realized she has a very sarcastic sense of humor. So we started presenting together with the goal of trying to make each other laugh. And when we do present, if you do have a chance to see Michelle and I together in person, we try to hide last minute photos or jokes for each other to make the other one laugh live and unexpectedly.

Jon Krohn: 00:21:37 That's funny. Yeah, she got some good giggles out of me while we were recording for sure. And also it's amazing given that, I mean I guess she has been now in the United States for a long time,

00:21:50 But something like sarcasm, I think of it as kind of one of the higher levels of comedy. I dunno if that shows my, maybe I'm not that sophisticated of a person, but sarcasm, I think it shows you have to have a pretty deep understanding of a concept to know that this is something that people are expecting. One thing, this is what usually happens when you use these kinds of words when you set up this kind of scenario. And so therefore people should be able to infer even if I dryly saying something that I'm making a joke. And so yeah, she lived in Korea until she was a teenager, and so to have that command, but it sounds like she gets a pretty high

command of, she speaks I think half a dozen different languages and

Amy Hodler: 00:22:33 A crazy number astounding to me that you can have that kind of command and not have it be native, but that's a very, yeah.

Jon Krohn: 00:22:44 Anyway, people can listen to episode 915 for more on Michelle. We will return to our regular programming now with graphs and so yeah. So is there anything else that you want to add in around graph algorithms or your graph algorithms or book before I move on to the next topic?

Amy Hodler: 00:23:01 Yeah, I would say graph algorithms. One of the reasons why I was smitten and graph algorithms are my favorite part of graphs in general is that if you imagine that either supply chain or fraud or crime or social network hairball, that when you have a lot of nodes in connections, you can have a very dense, complicated to air quotes look at and it's hard to see. If you look at tables row, if you look at an average, you're not getting the structure of a network. The topology of a network is really indicative of the behavior within a network, whether your network is growing, whether it's breaking apart, whether it's dying, whether it's clumping together, those are often driven by internal dynamics of your network science and you can't very easily see those with statistical methods or with machine learning methods. But graph algorithms compute over topology, they compute over structure, and so the results of them will actually tell you something that you didn't know about your network and you can infer meaning from that.

00:24:13 And that to me is so important right now because a lot of times I feel like in this moment, this machine learning gen AI moment, we are predicting the next item in a sequence, but we're not understanding our network, our

customers, our patients, our supply chain as well. And so understanding some inferring things about the network itself can be really helpful. Classic algorithm graph, algorithm everybody knows is page rank. So that was invented by Larry Page. We've all seen it when we Google and we know that that infers credibility of a source. That was the original intent. Now you can use page rank. You see page rank used for things like inferring lifespan of a telomere in a brain you see, which is a really cool use case. You see it for looking at toxicity of contaminants. It can be used so many different ways, but it's the reason is because it understands the topology of the network and then it tells you something like what is the important node in a network? Same with community detection and so many other graph algorithms.

Jon Krohn: 00:25:26 Wow. Those were a lot of great examples that you were able to just enumerate there. I love that that telomere one is particularly interesting to me. I have a PhD in neuroscience. Oh, fascinating. And so these kinds of things around, so some listeners may be aware that your lifespan as an organism is basically dictated by how long your telomeres are, which are things that are made by I think basically always your mom and your dad when they produced the sperm and or the egg that led to creating you. The telomeres were extended, basically it adds kind of garbage DNA or DNA that doesn't have any genes in it on the ends of your DNA, so that over your lifespan every time your cells replicate. So you go from a sperm and an egg to a single cell and then two and then four and then eight and 16.

00:26:30 And every time that happens, your DNA has to be copied. And then throughout your lifespan as your cells continue to divide, your telomere shrink, shrink, shrink because the mechanism that copies your DNA needs something to clamp onto. And so it clamps on right at the end. And so those telomeres that were added on by your parents

thank them for that is gradually eaten into over the course of your lifespan. And so a lot of what we consider to be aging is your telomeres being shortened, your telomeres run out in some of the tissues in your body and you're starting to eat into real meaningful genes and so interesting. So people who want to solve aging, a lot of it is about figuring out how you can have your telomeres grow while you're still alive without needing to create a sperm or an egg to be able to have those telomeres extend. The tricky thing is that telomeres extending is also one of the key causes of a metastatic cancer tumor.

Amy Hodler: 00:27:38 Yeah, it's definitely a balance. And the graph element of that is just understanding the connections between things and biology and biologists are very open to and understand a graph way of looking at things. So there, there's a lot of use cases in biology as well, probably not where the biggest funding is in graphs, but there's just a natural understanding when you talk to people with any kind of a healthcare life sciences background that they get the fact that things are connected and you can use graphs to understand the flow of resources, which you can model sugar, the flow of sugar through the body in that way. So anytime you're looking at highly complex systems, yeah, you end up getting into graphe theory behind that.

Jon Krohn: 00:28:31 You mentioned there that biology maybe isn't the place that a lot of resources are going for graph research. Where do you see a lot of resources going? If our listeners are thinking, wow, graph sounds super interesting or I already am a graph expert, I'd love to find lots of fertile ground, where do you think they could find it?

Amy Hodler: 00:28:48 It's changed over time. So I would say, I hate to say fraud, fraud and fraud, but finding bad actors, and when I say resources, I'm thinking commercial funding. And if you were doing a startup, what area would you focus in? So

fraud is always, I want to say evergreen, but it is crowded. So you have to have a different way of looking at it or some kind of unique quality of how you're looking at it. So always throw fraud in the mix. Whatever you're doing, you just throw it in there and it's a door opener and people understand it. A cybersecurity is hot again, it's starting to be a little crowded, but it's hot for grass because again, talking about how people try to hide and just the increase, the exponential increase in cyber crime right now means that people are looking at all different ways of tackling it.

00:29:46 And so I think grass and cybersecurity make a lot of sense and in particular when combined with other technologies as well. So it's another view on what you might already be doing. There's also some emerging use cases that I don't know if I would say reemerging that people had talked about previously but are now becoming more tractable because they're combining it with LLMs and gen ai. One of the ones that I would say are fairly low hanging fruit and probably people could build, I don't want to say build on their own, but for lack of a better word, you could probably build your own, is documentation analysis. So if you think about legal documentation, this is one area that I'm surprised not everybody's doing because it doesn't seem like that heavy of a lift, but legal documentation tends to have a lot of links to other caveats, other laws, sub clause, sub clauses of sub clauses.

00:30:47 And if you change this, what's the ripple effect of the connections to something else that you might actually be signing something that is impossible to fulfill because of the rippling connections and requirements. And so looking at legal documentation and helping your lawyers and your business analysts understand it better is easy. Low hanging fruit, it's easier now because you can use the LLMs to parse the language and then use the graph to

understand the connections between the different documents just for basic document review and understanding. We even see this in help desks, so help desk looking at their own documentation, their own complicated documentation. And I talked to somebody just in April that was trying to figure out their massive reams of documentation and help their support line get through their own documentation because there was just so many different connections and links and threads anytime you want to pull a thread, but also lawyers as well are looking at it. So those are some of the interesting use cases that aren't necessarily new, but because of gen AI coming in and large language models coming in, the combination are allowing these old new use cases to become just easier to accomplish.

Jon Krohn: 00:32:08 So those kinds of use cases that you were just describing there, like the legal document search is that graph rag,

Amy Hodler: 00:32:14 It can either be straight graph or you can use rag. I would say right now everybody's excited about rag, so I would expect that to be somewhere in the mix, but it doesn't have to be, and it wasn't always, so Caterpillar uses graphs and has for years for some of their support documentation and support staff, and that was before any of us had talked about rac, but I would imagine, I don't know, haven't talked to them lately, but I would imagine now they've got RAG in the mix and they probably have a chat bot that is talking to that documentation as well.

Jon Krohn: 00:32:51 So let's talk about that in more detail. I suppose if this is such an exciting area, graphic retrieval, augmented generation rag, probably a lot of our listeners out there are already familiar with the term, but it's this idea of being able to search over well. So you start off by having say a large number of legal documents or caterpillars in

		Caterpillar, that's like the, it's not a tractor company. What do you call those machines?
Amy Hodler:	00:33:16	I would call heavy equipment heavy because they do a lot of different equipment. Yeah,
Jon Krohn:	00:33:22	A lot of equipment with tracks on them,
Amy Hodler:	00:33:26	Hence the Caterpillar, right?
Jon Krohn:	00:33:29	Yes, yes, yes. Lots legs, lots of points on the ground I suppose.
	00:33:36	Yeah, so whatever these use cases, you have lots of documents. You start by converting all of those documents into a vector representation, meaning a location in some high dimensional space. So you can visually think about it like in a two dimensional space or a three dimensional space, but in practice it might have hundreds or even thousands of dimensions and the documents end up closer together because the meaning in the documents is more similar. And so LLMs have been great at figuring out that similarity and being able to get related documents close to each other. And so once you have all those documents close to each other, you can use rag retrieval, augmented generation to say, have a user ask a question. And that question is then in real time very quickly turned into that same kind of vector representation, the same kind of location in a high dimensional space. You can find related documents, you can pull them back and then you can use an LLM to search over that relatively small number of documents you pull back to generate some response to the user's query. So I mean you can correct or change anything that I just said about, but I'd love to hear now how we can take that rag idea and how it becomes rag.

Amy Hodler: 00:34:55 Yeah, so there are a number of reasons why RAG has some people, people feel like we are past the rag moment, just things get hot, people get disappointed, and then they want to improve it. But there are several reasons why RAG has plateaued. I would say a little bit in its performance, and that's why people are bringing in Graph as well to help out with it. It doesn't replace it augments, augments, rag. Anyhow, I don't know what the acronym of grag or who knows might be, but you've seen problems with RAG basically with things like getting a plateau when and your questions and answers don't align very well, especially semantically. And so you have this kind of meaning gap. So semantics just means meaning, but you have this meaning gap between what somebody's asking and the question and the context isn't really well understood. So you see things like that.

00:36:07 You see things like the ability to diversify the data you're pulling from. So RAG does really well with trying to extend what your retrieval is already doing. The retrieval effectiveness, pure semantic vector search can be insufficient. You might get an approximate match rather than an exact match, or you might want a match that has topological significance that you're just not getting. You might also need to do multi-step reasoning. So if you think about Agentic Rag where you need larger context windows or you need to pull in diverse data sets or you need to look at the previous responses to come up with a better response graph can help with that in general because again, we're looking at relationships and that gives us context and we're able to pull in topology as well.

00:37:15 I think of Graph Rag is Rag gives us a vector search, gives us really good summarization, it gives us really good fuzzy matches. If you're doing a full text search, you can get some exact matches, but graph gives us structure as well. Again, it's that topology and that gives us matches and answers of things like dependencies, serial things that

might have serial dependencies or you might have to do them in a sequence. It's also good at aggregations, especially if you're doing aggregations over multiple data sources and things like that. So I think of graph in the Graph Rag as Rag and it's like you don't want to just have one type of augmented generation, you want to include those other capabilities. Graph lets us do that. And it also helps us stitch things together because of the connections it has as well.

Jon Krohn: 00:38:14 So am I correct in understanding that they kind of happen in parallel that you would be doing a vector search as well as a graph search in parallel and so then you get the best of both worlds. You get the fuzziness of the vector search with the specificity of a graph search?

Amy Hodler: 00:38:31 Yes. Actually when you think about vector and graph in Graph Rag or it really should be called Hybrid Rag, that is the very thing you do see and the results that I've seen, and this is an evolving space, so ask me in six months question might be different or the answer might be different. But what I've seen is this hybrid in parallel that seems to get the best results and then you have a method to then blend the results at the end and then augment that in an ongoing fashion.

Jon Krohn: 00:39:03 Very cool. So I'm sure a lot of our listeners out there are saying very cool with me right now. And so all of our hands-on practitioner listeners, whether they're data scientists or software developers or AI engineers, they might be wondering how they should be getting started with graphs. It could be rag or just getting going on graphs In general, what are the key tools out there? Or I guess what are the ones that you recommend people get started with?

Amy Hodler: 00:39:30 Yeah, I always think thinking about your use case in mind, I mean graphs are just fun in and of themselves.

There's a lot of material. If you Google graph theory, there's a lot of fun material. But I would think about your use case and applying it in that manner. I try not to recommend specific tools, but there are a couple of vendors that have really nice one-on-one material that you can use to get up to speed. I would say right now, don't lock yourself into one methodology. In the past, several years ago, maybe three, four years ago, it was assumed if you were going to add graph to your capabilities, you had to have a graph database that is no longer the case. There are several vendors out there that allow you to project graph with a computational graph, so classic graph engine or a graph layer.

00:40:24 There are several vendors that also allow you to poke into your table data and ask a graphic question of your table data and just use that projection to answer the question and then drop it. And so depending on your need, there's a lot of different options for you and there's no one option that's good for every situation. So looking at your own tech stack is probably real important. There's a couple conferences I really like. If anybody to give a shout out to the open data science conference group, they do conferences in the East coast, west Coast, Europe. They also do virtual conferences or virtual training as well. I like that because you get a broader spectrum of opinions and different ways of looking at graph, but also how it fits into the bigger picture as well.

Jon Krohn: 00:41:23 ODSC is my favorite conference as well. And I'm not just saying that because they sponsor the show and they literally, actually in your episode, Amy, you can possibly know this, you don't know what ads we're going to put in.

Amy Hodler: 00:41:35 No.

Jon Krohn: 00:41:35 But at around the 30 minute mark in this very episode, there's an ODSC West 2025 sponsor message. I

absolutely love ODSC. I think it's the best for hands-on practitioners

Amy Hodler: 00:41:50 Would agree

Jon Krohn: 00:41:51 Because I'm in New York, I have the privilege of speaking at East in Boston in the spring most years, and I get out to West. It's pretty much always on Halloween. I get out there whenever I can. Sometimes my travel schedule doesn't permit it, but if it does, I'm always at West as well. I love it.

Amy Hodler: 00:42:13 Well, Michelle and I will be there this year.

Jon Krohn: 00:42:17 I'll have to try to make sure I get there.

Amy Hodler: 00:42:20 We could do something fun. Yeah,

Jon Krohn: 00:42:22 We could do, we could do standup comedy. You should just stick to technical stuff.

Amy Hodler: 00:42:30 Terrifying me. I don't think, I'm never funny on purpose, but I'm often funny. So there you go.

Jon Krohn: 00:42:37 Actually, I recently was on another podcast called the Modern CTO podcast, which is a cool show and the host is very funny and he actually, he now spends other than, so he professionally hosts the modern CTO podcast. That's his main job, but a huge amount of his time approaching a full-time job is being a standup comedian now as well. So Joel Beasley, and it's pretty interesting to hear, I think by the time your episode is out, my appearance on the modern CTO podcast should be out. So there's cool things that he talks about in the episode around using data and analytics to and using S to review all of his standup routines. And so he's trying to get a particular number of laughs per minute from the audience and he's trying to match, so he's done analytics

around what other professional standup comedians, if you get a Netflix special, how many laughs do you get per minute on average? And so he's trying to work his way up to that.

Amy Hodler: 00:43:45 Wow. Terribly. That reminds me of a story. I had somebody once asked if I would come to their offsite, they were quants and they're like, can you come to our offsite talk about graphs and be funny? I was like, you're terrified me. I don't. Oh, okay, we will See.

Jon Krohn: 00:44:05 That's a lot of pressure. A lot of pressure when you feel like you have to be. So yeah. So you were very diplomatic about your answer to the tools, which I greatly appreciate. And I suspect that that's somewhat related to you being founder and executive director of GraphGeeks. Tell us about GraphGeeks, how it got started and why someone should reach out you if they need support.

Amy Hodler: 00:44:33 Yeah, so GraphGeeks, I love to tell people that I started it because I got lonely, but that is actually part of the truth. So I'd been in the graph space for several years, I dunno, over 10 I guess. And I have been at several different vendors. And when you're working for a vendor, it's wonderful to give you focus, but you also have that focus also gives you a particular way of looking at the graph's landscape and world. And so when I left the last graph company I was at just had a lot of wonderful conversations with people that were either starting things up or had a lesser known, lesser funded approach to graphs and really wanted to highlight that. And the other thing is I was just missing my graph friends. So the interesting thing about graph folks, they're often a small subset at a company.

00:45:27 You usually don't have a hundred person graph team. You sometimes have a one or five person graph team. And so we hold onto each other pretty closely. We talk to each

other just naturally. Well, we look at the world through relationships. So we naturally keep our relationships, we help each other out over the years and we stay in contact, we go to the same conferences. So it's a bit of a tight-knit community. And so when I left the vendor space, I immediately missed my community and how we looked at things and I was looking around and I realized there was no vendor agnostic graph community online it that I could easily reach out with and interact with on almost a daily basis if I wanted to. So I thought, what the hell, I can start this. If I start this and I fail, nobody can fire me.

00:46:15 Why not? And so I started GraphGeeks. We are a GraphGeeks.org. If people are looking for us, there is a discord, sizable discord community that helps each other out. So we have a I need help section, we have resources, probably not as many training resources that I would like to get right now, but there are a lot of graph practitioners. So if you're stuck in the middle of a really gnarly graph problem and you're looking for help, you can go on the Discord channel and say, Hey, I'm stuck. What have you seen? If you've read a paper, love it, hate it, don't understand it. Reach out. If there's somebody you would like to see on the GraphGeeks either webinar, podcasts, what have you, I've got a YouTube channel, let me know. I try to get that information out there and help people with things on what is the difference between RDF and property graph. We've got a couple hours of that material out there. How do you design to optimize query? We've got some stuff on that. So there's a lot of interesting topic, but the well is deep. And so over time we'll just be adding more and more material. I also have many volunteers and opportunities for volunteers. So if people just wanted to geek out, we do that. So that's how I got started.

Jon Krohn: 00:47:42 Nice work. Amy. It sounds like you're doing an amazing thing for the graph community. You are now an invaluable node in that network,

Amy Hodler: 00:47:49 A pivotal node, call me a pivotal

Jon Krohn: 00:47:51 Node, A pivotal node. Thank you. So before I let you go, one last technical question that I want to get some insight from you on is what is changing in graphs? What's next? So we've spent this episode learning about why graphs are cool, what they're useful for. You gave us some direction on tools that we could be graphing. And so yeah, what's next? Some of the things that you mentioned to me before we started recording included multimodal included graphs for LLM memory and causal graphs. Maybe we could touch on each of those quickly.

Amy Hodler: 00:48:29 Yeah, so I'll quickly go through the major changes. One is that I already discussed a little bit is framework diversity. So are the query engines are getting better. So you don't have to have a database, different types of graph databases are becoming available. You also have hyperscalers that are getting into reentering the graph space. So lots of choices on framework. So that's a big one. Multimodal I would put there, I would put out, well maybe I should say graphs and AI and what bringing them together is allowing from a use case standpoint, we talked a bit about that. And then multimodal, which is being able to graph different types of data. So one of the things a colleague of mine, David Hughes, shout out to him and I do present on is this idea of modeling an image as a graph. And so most of the time we talk about graphs, people think about lexical graphs, so graphs of words or graphs of concepts, those are the traditional uses.

00:49:38 However, you can graph an image. So if we have a picture of me holding my coffee cup, you have the main images is Amy, but there's a coffee cup in front of me to the right, and that relationship has meaning as well. And so being able to connect those as meaning allows us to do things. If we're looking at, for example, and we've done this looking at a fleet of ships and some are ahead of the

other, and you can graph that relationship. And then if you look at that relationship over time, you can also estimate the speed. Are those ships coming together? Are they pulling apart? Do they look like they might be antagonistic to each other? So there's all of these things that you can do with different data types. So again, moving to images, we've also added in audio to that. And so for example, we did that with police cars where you hear them in a video frame but you don't see them.

00:50:40 And with Doppler effect, you can tell what direction the police cars are heading and you can do that by graphing it. And to me that's exciting, not just from a graph rag standpoint, which is what most people want to talk about, how do I use that with my graph rag? But just this idea of something we have done with graphs forever, which is modeling the relationships between things. We haven't extended it to things in a image or things in audio. And to me that just opens up to all sorts of other use cases like detecting things in sonar to, again, directional speed in an image to understand a grouping in an image of people. Is there a relationship that we can infer based on how people are standing next to each other? So there's that to me. Sorry, multimodal, very, very fascinating area, really cool.

00:51:37 But the other one or the other two that I would be remiss if I do not mention them first is graph as memory. So graph provides us a way to capture context and context is really important for ai. And so if you think about the context windows of an agent, they're relatively short right now. So there's a couple of really interesting papers, Zep, which I have sitting on my desk right now. Temporal knowledge graph architecture for agent memory, a must read if you're interested in extending agent memory and then mem zero building production ready AI agents with scalable long-term memory. Those two papers really significant I think in looking at how you use the context

saving ability of a graph to store memory for agents either for just very simply extending the context window and you can basically store a context and then retrieve it later when you need it or even longer memory. So going beyond a typical context window that I think is going to be super hot by the end of the year. If you're into graphs and you haven't thought about graphs as memory for agents, take a look because that's something that I think in six months or less people are going to be talking about.

Jon Krohn: 00:53:03 It is something that's been on my radar as well, that mem zero paper is something that keeps coming up. So yeah, I agree. Keep an eye on that.

Amy Hodler: 00:53:12 The final thing as far as emerging if we want to add this at some point is causal graphs. So understanding not just the prediction, being able to predict that something happened, but why it happened, I think is something that in two-ish years will be the next hot thing because we are doing a really good job of predicting the next item in a sequence, but we don't know why that sequence happens. And graphs because you have those two nodes, those circles with an arrow between it, those arrows or those links can be an arrow. And so you can link things in a sequence and you can try to understand influential cause in whether you're talking about economics or biology or what have you. So that's the other area to look out for, but probably a little further ahead.

Jon Krohn: 00:54:02 Nice. Thank you for that. Look into the future, even though things move quickly, it does feel like your insights looking out six, 12 months, maybe even longer are going to end up being really, really helpful. Amy, this episode has been incredible. You're an absolute pro. It's unsurprising that you host your own podcast, that you do all of this public facing graph knowledge distribution work because people wouldn't be able to tell this because we edit all the episodes, but everything in this episode

was done in a single take, and so there's been no retakes, there's been no umming and aing thinking about answers. Amy's just done everything off the cuff flawlessly. It's been a joy recording with you. Before I let you go, I ask all of my guests for a book recommendation. Do you have one for us?

Amy Hodler: 00:54:55 Yes. Actually I joined a book club, which I highly recommend joining an old fashioned book club. It's really wonderful to talk to humans about what they are thinking. The book is called These Strange New Minds, and it is about LLMs and thinking about them from everything from a technical standpoint to a philosophical standpoint. Are they thinking? Are they not thinking? Do they have the appearance of thinking? What does thinking mean and what do we want LLMs to say? How do we look at optimizing and regulating? So I think in this moment in time that we are all in, understanding LLMs a little deeper and from a couple different viewpoints is just a huge benefit and I highly recommend the book. It's exceptionally well written.

Jon Krohn: 00:55:46 So it's nonfiction. I mean it's not like a fictionalized. Yeah,

Amy Hodler: 00:55:51 No, no, it is not. It is a multidimensional look at large language models and what it means for models to actually speak back to us.

Jon Krohn: 00:56:01 And so this book club that you're in, is this the typical kind of book that you read? Are other people in your book club also kind of data AI people like You are what's going on here?

Amy Hodler: 00:56:12 Yeah. Yes, it is a typical book. We also did one on responsible AI called AI Snake Oil. That's another really good one to look at, but definitely has its opinion so far. Yeah, we've been picking data and AI topics and it's a group of us that are in the tech industry, but it's even

within the tech industry. These concepts are so large, having people with different viewpoints. We have somebody from a security background, we have somebody with a responsible AI background, an ethics background. We have people that are just building product. So I think having diverse people and mindsets, people from different countries as well really shows that even when we have our own opinions on something, there's probably a different way we haven't looked at it. So anyhow, join a book club. Doesn't matter what it's for.

Jon Krohn: 00:57:09 It's a great idea. It's something that I, it's on my list of things to do. I hope to get there. I even just reading more. It's something that I fantasize, just like my fantasy about being under leaves. I also fantasize about being under the leaves of a book. Yeah,

Amy Hodler: 00:57:31 We can talk about that another time. I do think about doing more. It has been very rewarding and it's very useful.

Jon Krohn: 00:57:38 That's why we do these book recommendations at the end of every episode. I know how valuable it is, even if I would love to read every single book recommendation that I get from my guests, and there's just no possible way I can do it. Fantastic, Amy. So yeah, as I already said a few minutes ago, this has been an amazing episode. You've been an extraordinary communicator. We've already talked about GraphGeeks, which people can visit@graphgeeks.org. Where else should people follow you after this episode to get more of your brilliant thoughts?

Amy Hodler: 00:58:04 Well, they can also follow me on LinkedIn. I'm very easy to find. I have an unusual last name, so Amy Hodler, H-O-D-L-E-R. Easy to find me there. That's primarily where I post that in. I'm on GraphGeeks.

Jon Krohn:	00:58:18	Fantastic. Thank you so much, Amy. It has been such a joy to have you on the show. And yeah, hopefully we can get you on again in a few years and see how the graph world has come along.
Amy Hodler:	00:58:28	Absolutely. Thank you, Jon. It's been a real pleasure and I've had a lot of fun, and I appreciate all your really insightful questions.
Jon Krohn:	00:58:38	Wow. Another fun and informative episode today. In it, Amy Hobbler covered how graphs capture relationships and data using nodes and edges with properties that include quantities, strengths, and various data types. How graph algorithms like PageRank compute over network topology to reveal insights about structure and behavior that traditional statistical methods miss, making them powerful for applications like fraud detection, supply chain optimization, and recommendation systems. She talked about Graph Rag and how it enhances traditional retrieval augmented generation by adding structural context and topology to semantic vector search. And she talked about exciting emerging graph applications including multimodal graphs, graphs as memory systems for AI agents and causal graphs.
	00:59:22	As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Amy's social media profiles, as well as my at superdatascience.com/923. Thanks to everyone on the SuperDataScience podcast team are podcast manager, Sonja Brajovic, media editor, Mario Pombo, partnerships manager, Natalie Ziajski, researcher Serg Masís, writer Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing yet another excellent episode for us today for enabling that super team to create this free podcast for you. I encourage you to check out our sponsors. You can support this show by clicking

on our sponsors links in the show notes. And if you are interested in sponsoring an episode yourself, you can get the details on how at jonkrohn.com/podcast. Otherwise, share this episode with folks who'd love it. Review the episode on your favorite podcasting platform or YouTube or wherever you consume podcasts. Subscribe if you're not a subscriber, but most importantly, just keep on listening. I'm so grateful to have you listening, and I hope I can continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.