# Scrape and graph your way to conference glory: Building the ultimate Call for Papers tool

**Lju Lazarevic**
Developer Advocate
@ellazal
http://lju.io

**This story begins with a Developer Advocate, facing a time-honoured challenge…**

***What conference should I submit to?***

# I’m a Developer Advocate at Neo4j

What does a Developer Advocate do?

- Think about problems a developer is trying to solve
- Come up with relevant examples
- Help you with the tools they love!
  - (I really, really love graph databases)

Ways a Developer Advocate might do this:

- Help out in the community
- Write blog posts and create example code
- **Present at conferences and meetups**

# Challenges behind finding suitable conferences

- No centralized source
- No obvious way to search for certain topics/tags/keywords
- Sometimes there are no tags/topics/keywords!
- “Conference Driven Development”

**This particular Developer Advocate was looking for potential conferences a few months ago…**

***There's got to be a better way...***

# Then I had my 💡 moment...

- I did a bit of web scraping many years ago
- I am graph database geek - all about those connections!
- What if I could bring these approaches together, and:
  - Create a Conference Call for Papers tool!



# And there’s more!

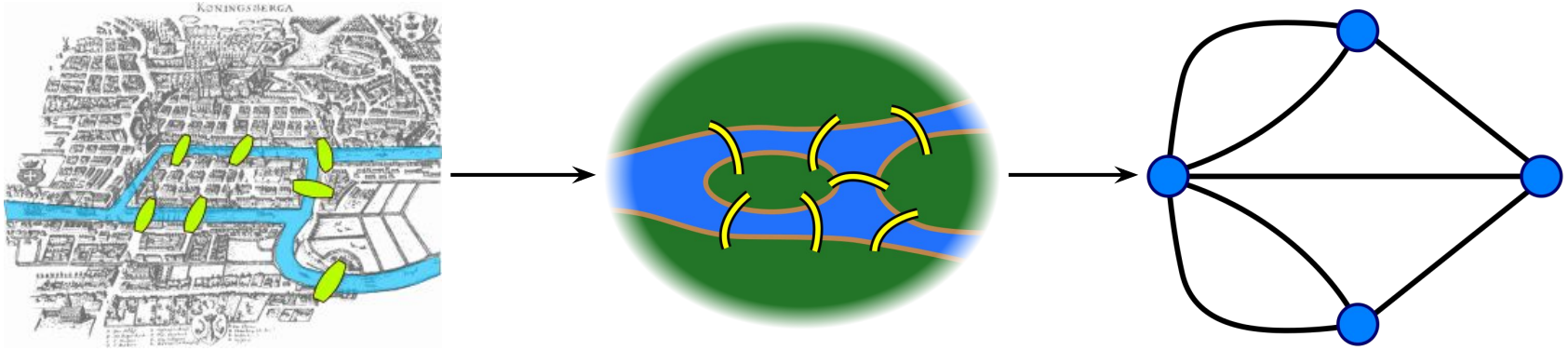
We could use this tool to:

- Get insight into tech trends
- Group similar-themed conferences together
  - Maximise that “Conference Driven Development” effort!
- Build more relevant examples for the developer community
- And so much more!

# But first! What is a graph database?

# A graph is...

...a set of discrete objects, each of which has some set of relationships with the other objects




*Seven Bridges of Konigsberg problem. Leonhard Euler, 1735*

# A graph database is...

...a database that stores data entities and their relationships in a graph structure.
Anatomy of a (property) graph database includes:

**Node** (Vertex)

- Main element from which graphs are constructed

**Relationship** (Edge)

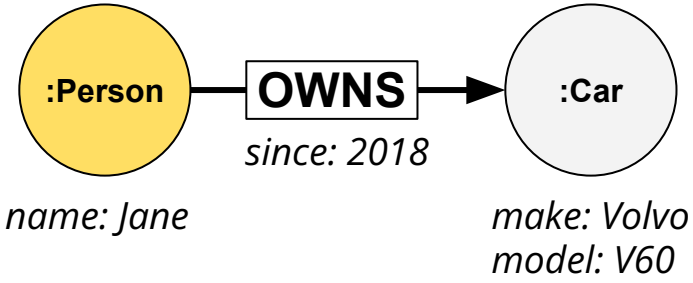
- A link between two nodes, has direction and type

**Label**

- Define node category

**Properties**

- Enrich a node or relationship



# What are the differences between a relational database and a native (property) graph database?

# Relational databases - joins on read

- Typically normalized data & mapping tables
- Joins at query time to reconstitute data and find connections
- Hypothesize on ‘is the data connected?’
- Increase in joins → exponential increase in query execution time

**Are** Ann, Dan and Jan connected?

| person_id | name |
|---|---|
| 1 | Ann |
| 2 | Dan |
| 3 | Jan |

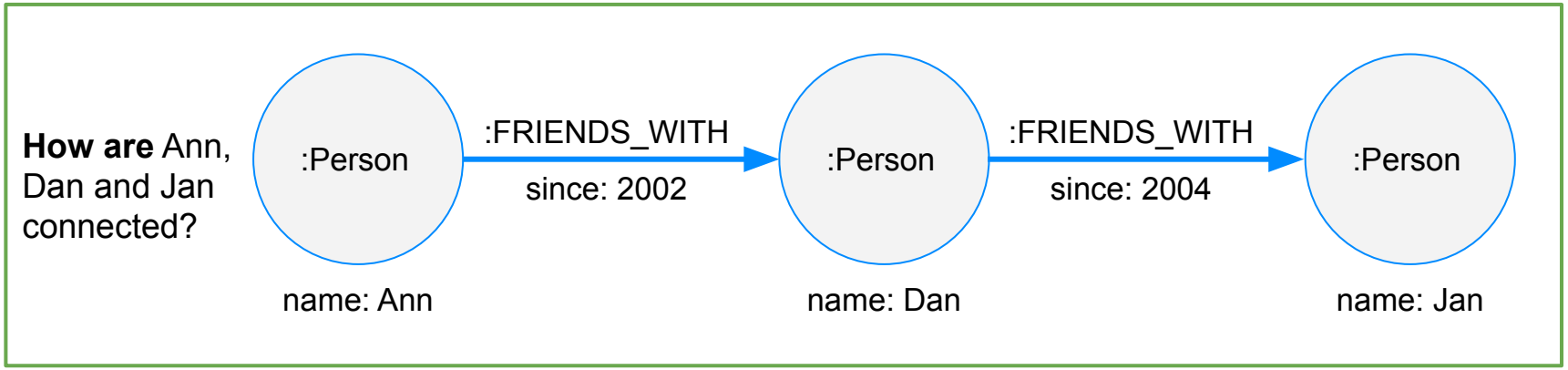
| friend1_id | friend2_id | since |
|---|---|---|
| 1 | 2 | 2002 |
| 2 | 3 | 2004 |



neo4j

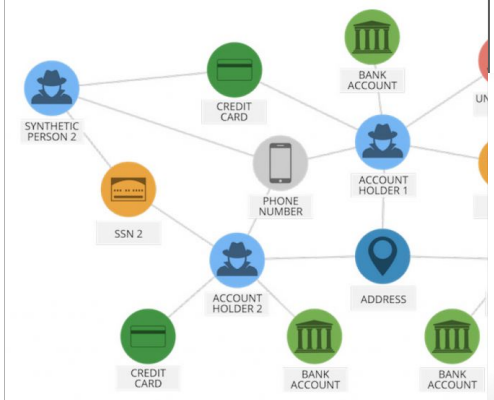
# Native graph database - joins on write

- Relationships are first-class citizens
- Joins at write time - “physically” join connected entities
- Hypothesize on ‘how/why is the data connected?’
- Increase in joins → linear increase in query execution time

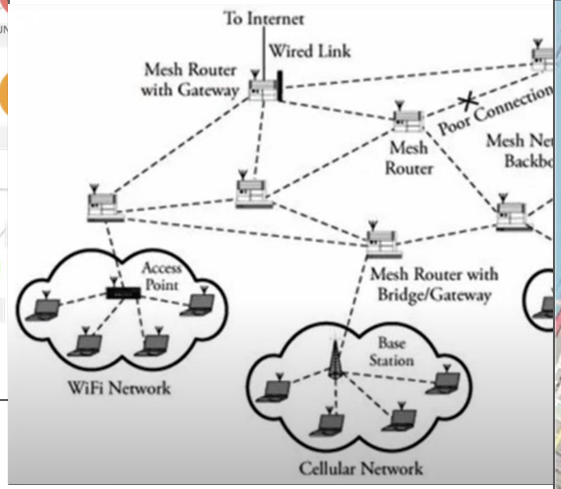


# Great graph use-cases

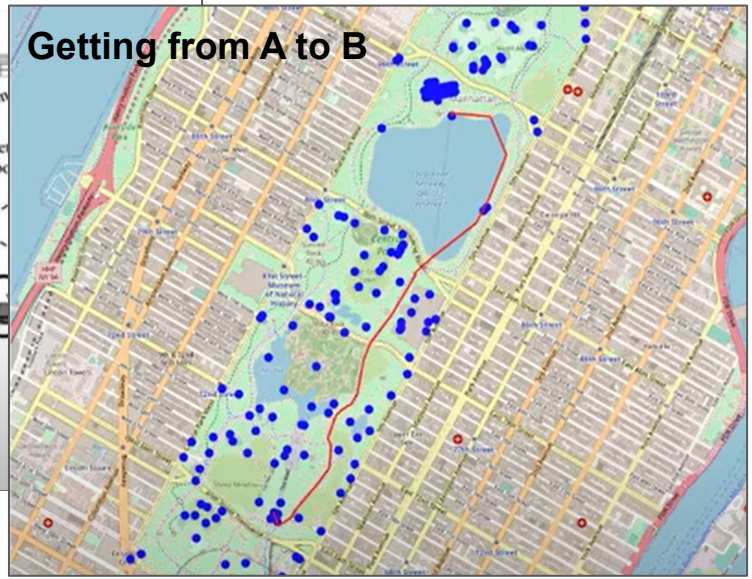
## Detecting Fraud




## Predicting outages




## Getting from A to B



# So, how do we build this CfP tool?

# Building the CfP tool

1. Identify common places for Call for Papers information
    a. Web search
    b. Ask around
    c. Past experience

PaperCall

Home | Event Directory | Past Events | Pricing

# The easy way to manage your call for papers

PaperCall enables event organizers to easily manage their call for papers and talk submissions. Speakers are already using PaperCall to manage are you ready to hear from them?

Start now ›

sessionize

Features | Pricing | FAQs | Testimonials | Contact | Login | Try For Free

# The smart way to manage Call for Papers, Speakers and Agenda for your event.

Cloud based, safe and easy. Your speakers will love it, too!

EVENTS | TALKS | COMMUNITY | PRICING | CREATE EVENT | LOGIN | REGISTER

# Events for Tech People

Particip technical events anywhere in the world

START TODAY

## FEATURED EVENTS


SUN, 05 JUN 2022 9:00 AM

Agile Testing Days


THU, 04 NOV 2021 9:00 AM

Code BEAM


MON, 25 OCT 2021 9:00 AM

International PHP


MON, 08 NOV 2021 9:00 AM

GOTO


MON, 08 NOV 2021 8:00 AM

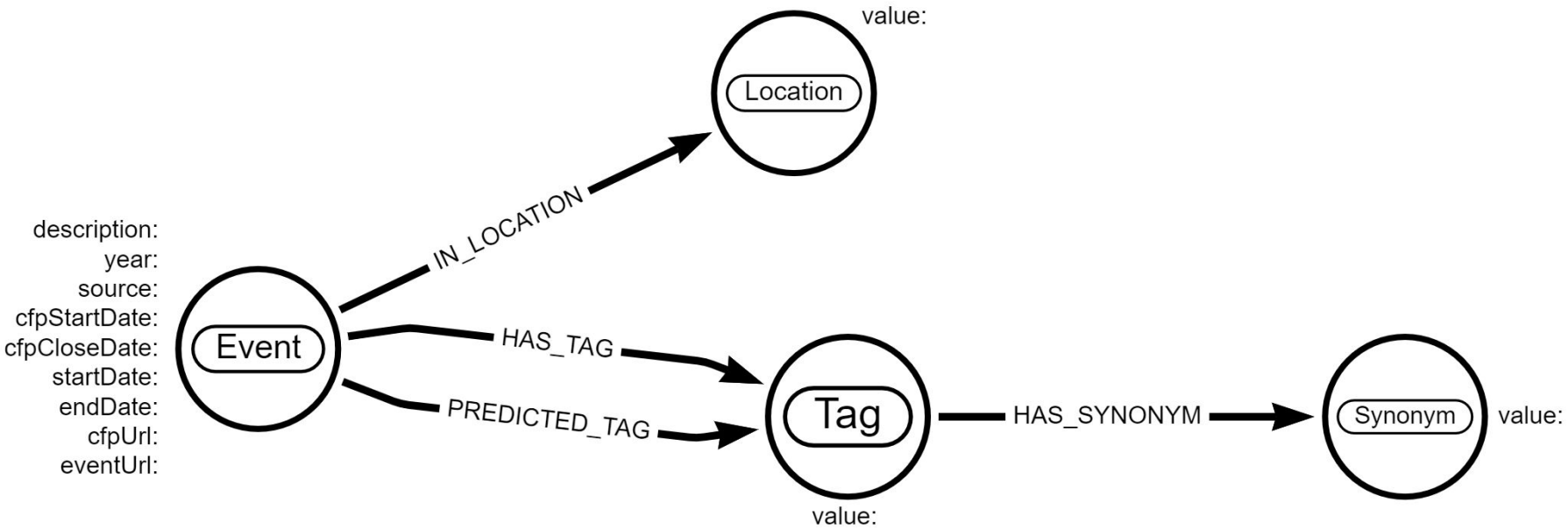
W-JAX 2021

# Building the CfP tool

1. Identify common places for Call for Papers information
   a. Web search
   b. Ask around
   c. Past experience
2. Identify what data do we need and build a data model

# The data model

Drawn using Arrows: http://arrows.app



# Building the CfP tool

1. Identify common places for Call for Papers information
   a. Web search
   b. Ask around
   c. Past experience
2. Identify what data do we need and build a data model
3. Figure out how to get at individual Call for Papers information
   a. Is there a directory we can iterate over?
   b. Is there a site map we can use?
   c. Can we carefully craft a search engine query?
4. Examine the CfP source code to figure out how to extract the data
5. Extract the data and load into the database, based on the data model

This XML file does not appear to have any style information associated with it. The


```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  ▼<url>
    <loc>https://sessionize.com/devfest-minna-2021/</loc>
    <priority>0.5</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/data-toboggan---world-azure-synapse-day/</
    <priority>0.8</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/ee0264017cf2444792df9c5a19f2eea3/</loc>
    <priority>0.8</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/global-ai-back-together---cleveland-2021/<
    <priority>0.8</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/data-sat
    <priority>0.8</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/global-a
    <priority>0.8</priority>
  </url>
  ▼<url>
    <loc>https://sessionize.com/global-a
    <priority>0.5</priority>
```


PaperCall Home Event Directory Past Events Pricing Help Login

## Event Directory

All Events Open CFPs Only Only Recurring Events/Meetups

Add Your Event! Create Your CFP!

Offers Travel Assistance

Save Search as an Alert

Search

Tag search only: tags: first tag, second tag, etc...

PRO EVENT **INTENT - The security research summit. - Virtual**

INTENT http://www.securityresearchsummit.com/

**Event Dates:** November 16, 2021

**CFP closes at** October 04, 2021 19:59 UTC

October 04, 2021 20:59 BST (Local)

Submit Now! I'm Attending! Add to Calendar

Submit Now! I'm Attending! Add to Calendar

Submit Now! I'm Attending!

Help

EVENTS TALKS COMMUNITY PRICING CREATE EVENT LOGIN REGISTER

## Find Tech Events

UPCOMING CALL FOR PROPOSAL ONLINE PAST

Search for events Enter location All times SEARCH

14 DAYS LEFT

SUN, 26 SEP 2021 9:00 AM

Matillion for Snowflake Training & Online Certification Course

Online

MON, 11 OCT 2021 1:00 PM

Color Psychology for Behavioral & UX Design (Fall 2021)

Online

CONFERENCE

September 14

www.greece.affiliate.events

WED, 13 OCT 2021 10:00 AM

Greece Gambling Conference 2021

NJV Athens Plaza

curb

view-source:https://www.papercall.io/djangocon-us-2021

```
  <div class="justifize__box">
    <div class="subheader__logo">
      <img class="" src="https://papercallio-production.s3.amazonaws.com/uploads/event/logo/4267/mid_300_DjangoCon-2021.jpg" alt="Mid 300 djangocon 2021" />
    </div>
    <div class="subheader__group">
      <h1 class="subheader__title">DjangoCon US 2021</h1>
      <h1 class="subheader__subtitle">
        Online
          October 21, 2021, October 22, 2021, October 23, 2021
      </h1>
        <a target="_blank" href="https://2021.djangocon.us/">https://2021.djangocon.us/</a><br>
        <span>Tags: <a href="/events?keywords=tags%3A+Python">Python</a>, <a href="/events?keywords=tags%3A+Django">Django</a>, <a href="/events?keywords=tags%3A+documentation">Documentation<
    </div>

  </div>
    <div class="justifize__box pull-right">
      <h1 class="subheader__subtitle">
        <a href="https://www.facebook.com/dialog/feed?
  app_id=929684357128596
  &amp;display=page&amp;caption=PaperCall.io
  &amp;name=DjangoCon US 2021
  &amp;picture=https://www.papercall.io/assets/logo-papercall.svg
  &amp;link=https://www.papercall.io/djangocon-us-2021
  &amp;redirect_uri=https://www.papercall.io/djangocon-us-2021"><i class="fa fa-facebook" data-toggle="tooltip" data-placement="bottom" title="Share on Facebook"></i></a>
        <a href="https://twitter.com/intent/tweet?text=Submit to the DjangoCon US 2021CFP! https://www.papercall.io/djangocon-us-2021"><i class="fa fa-twitter " data-toggle="tooltip" data-pla


        <button class="fa fa-external-link copy-to-clipboard" data-clipboard-text="https://www.papercall.io/djangocon-us-2021" data-toggle="tooltip" data-placement="bottom" title="Copy to Cli
      </h1>
    </div>

  </div>
</div>
</div>

<div class="container">
  <div id="flash_notices">

<div class="row">
  <div class="col-md-12">
```

# Success! We're finished!

Not quite… We have some issues to deal with:

- Not all of the conference platforms that we looked at have tags
  - Use the tags we have
  - Scan across event descriptions and titles
- There are some data quality challenges
  - We'll look at some small fixes now
  - Explore options to sort in the next iteration

# Are we done now?

Almost! We don’t have many tags:

- Good source of technology tags - StackOverflow
- Use the StackExchange data explorer to pull top 200 tags
  - Also synonyms
- Some fuzzy matching after

Yet (more) issues:

- Not all tags make sense
- Use some basic stats to get rid of them
  - Proportion of talks with that tag
  - Frequency of tag appearing across titles/descriptions

Still no perfect, but good enough!

https://github.com/lju_lazarevic/cfptool

# The tools in use

- Lots of googling
- Jupyter notebook
- Beautiful Soup & Google search
- StackExchange data explorer
- Neo4j stack
  - Sandbox - No-download trial database
  - Python driver - API to connect to the database
  - APOC library - a collection helper functions and procedures
  - Browser - developer aide for queries and visualisation
  - GDS library - graph algorithms for data science

# We’re there! Let’s find some conferences to submit to!

# Let’s go find some conferences!

- What are the most ‘popular’ tags?
- What CfPs are closing within the next month?
- What CfPs have the theme of data science/machine learning?
- What CfPs have tags of data science and python?
- What conferences can be grouped together by similarity?

# So what’s next?

# What’s next?

- General code and data clean-up
- More data!
- Find other sources for more tags and synonyms
- Use of NLP to extract more meaningful tags/exclude meaningless tags
- Create conference ‘themes’, e.g. DevOps, Data Science, etc.
- Load up historical conferences
- Explore options for importing historical talk titles

Repo: https://github.com/lju-lazarevic/cfptool

# Want to learn more about graphs?

**Free online training and certification:**

- *dev.neo4j.com/learn*

**How to, best practices, hands on and community stories:**

- *dev.neo4j.com/videos*

**Come say hello :)**

- *dev.neo4j.com/chat*
- *dev.neo4j.com/forum*