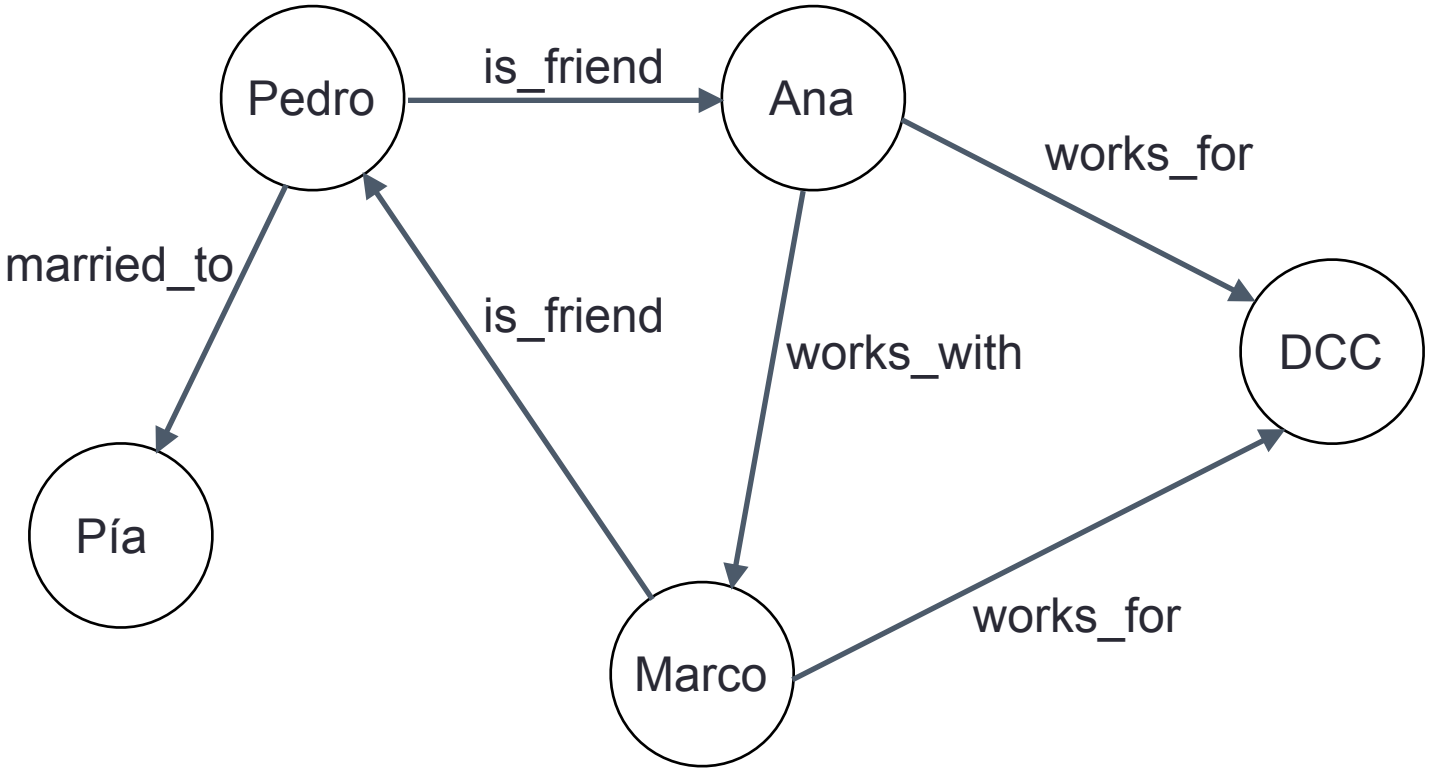
Consultas de Conectividad en Bases de Datos de grafos

Juan L. Reutter

DCC PUC-Chile

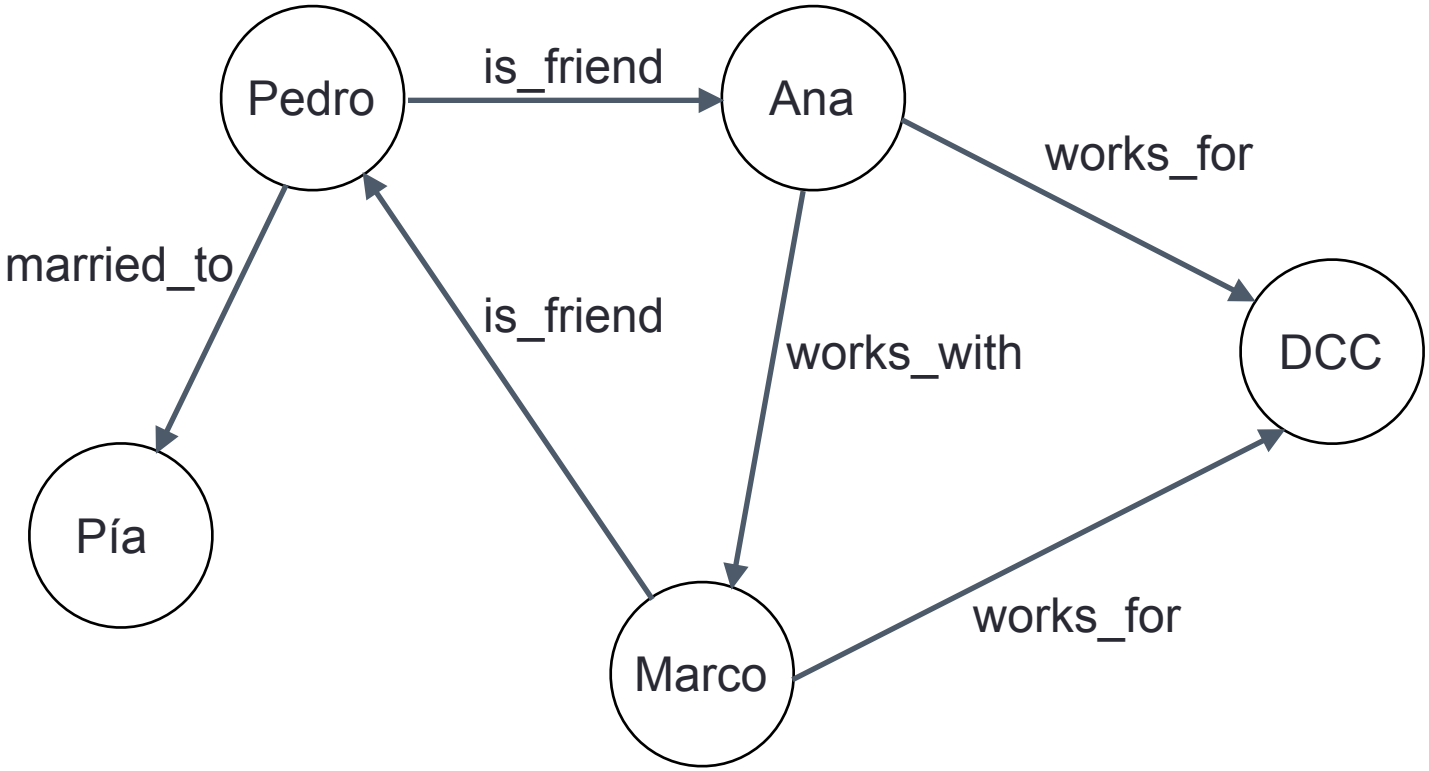
Center for **Semantic Web** Research

Bases de Datos de Grafos

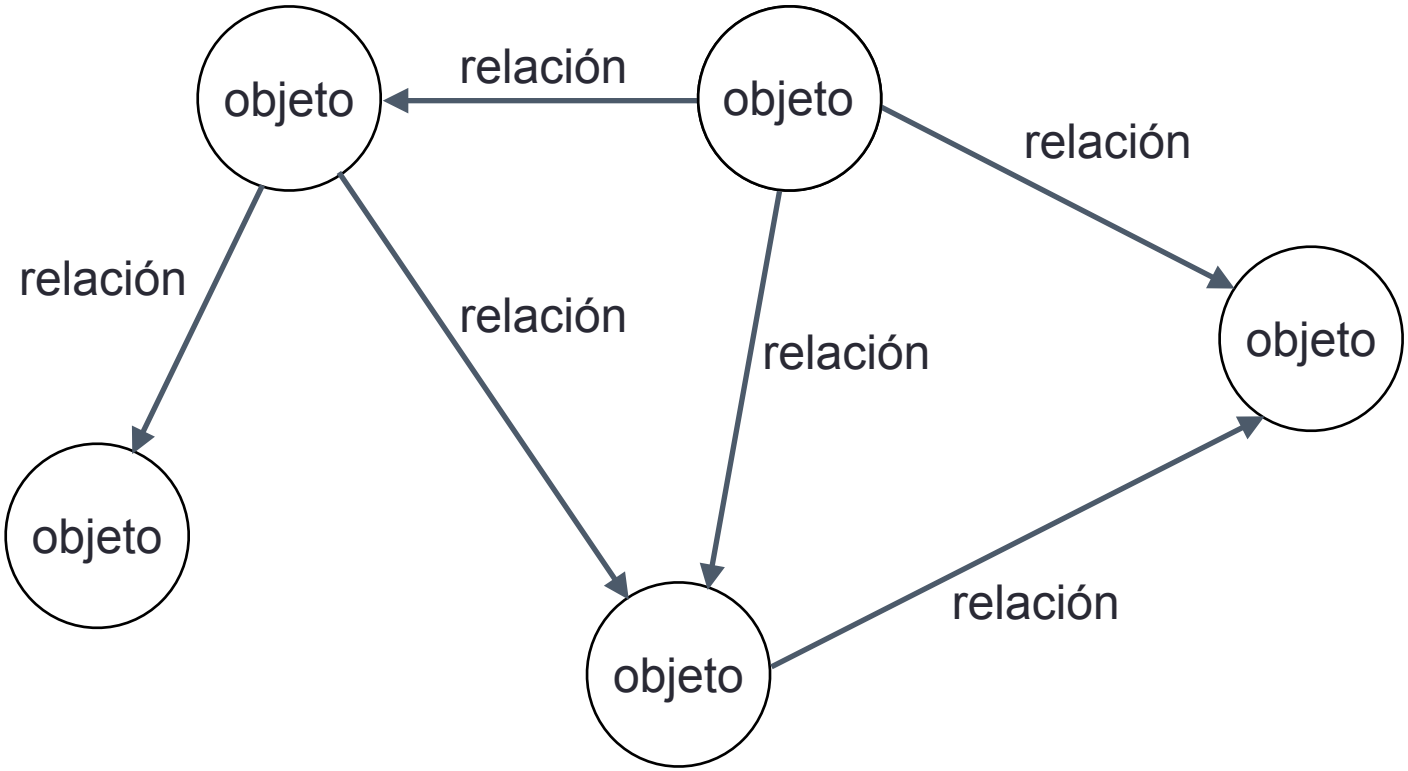


Graph Databases

- Redes Sociales
- Bases de Datos Biológicas
- Modelos Geográficos
- …



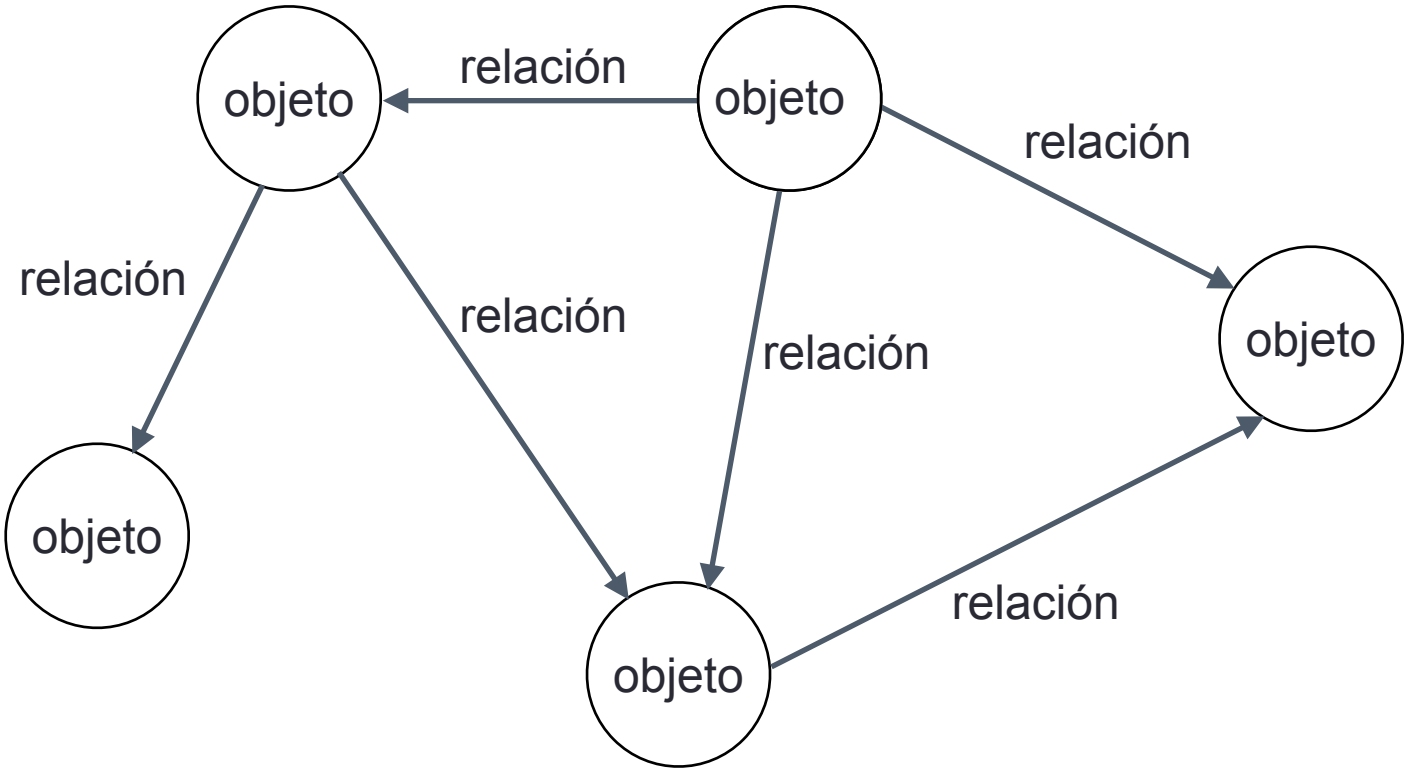
Graph Databases



Graph Databases

Modelo Básico:

- Nodos Representan Objetos
- Aristas son relaciones

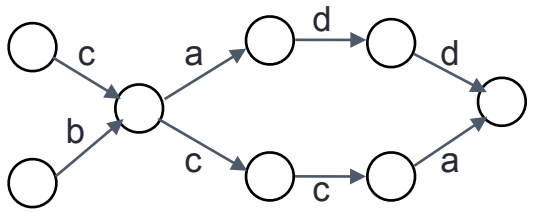


Consultas clásicas: Pattern Matching

Consultas clásicas: Pattern Matching

Consultas clásicas: Pattern Matching

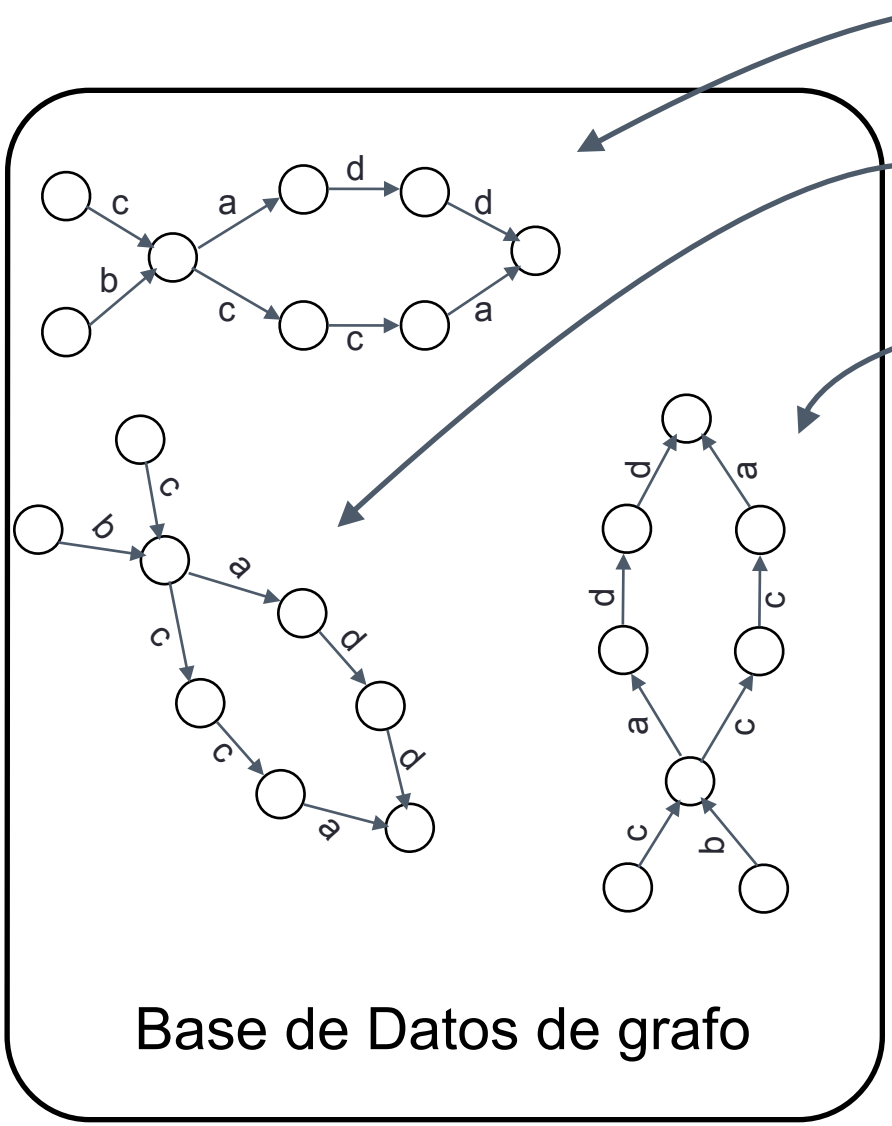
Base de Datos de grafo

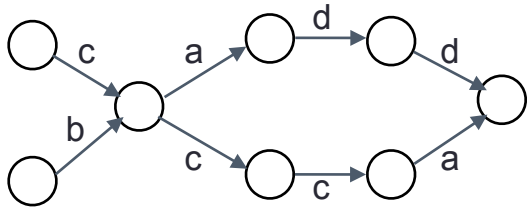


Patrón pequeño de interés

- Encontrarlo en un grafo

Consultas clásicas: Pattern Matching





Patrón pequeño de interés

- Encontrarlo en un grafo

El problema de pattern matching corresponde al clásico problema de Isomorfismo de subgrafos

- Este problema es NP-completo
- Aproximaciones y heurísticas para resolverlo
- Campo activo en computación

El problema es que muchas aplicaciones necesitan más que pattern matching

Consultas de conectividad:

- LinkedIn: ¿Puedo contactar a un abogado experto en recursos naturales?
- Criminología: ¿Quienes son todas las personas con contactos directos o indirectos con un sospechoso?
- Logs de Wikipedia: última versión del artículo donde no participó cierto usario

El estudio de estas consultas es un campo de investigación activo en Ciencia de la Computación

Consultas de Conectividad en Bases de Datos de grafos

Juan L. Reutter

DCC PUC-Chile

Center for **Semantic Web** Research

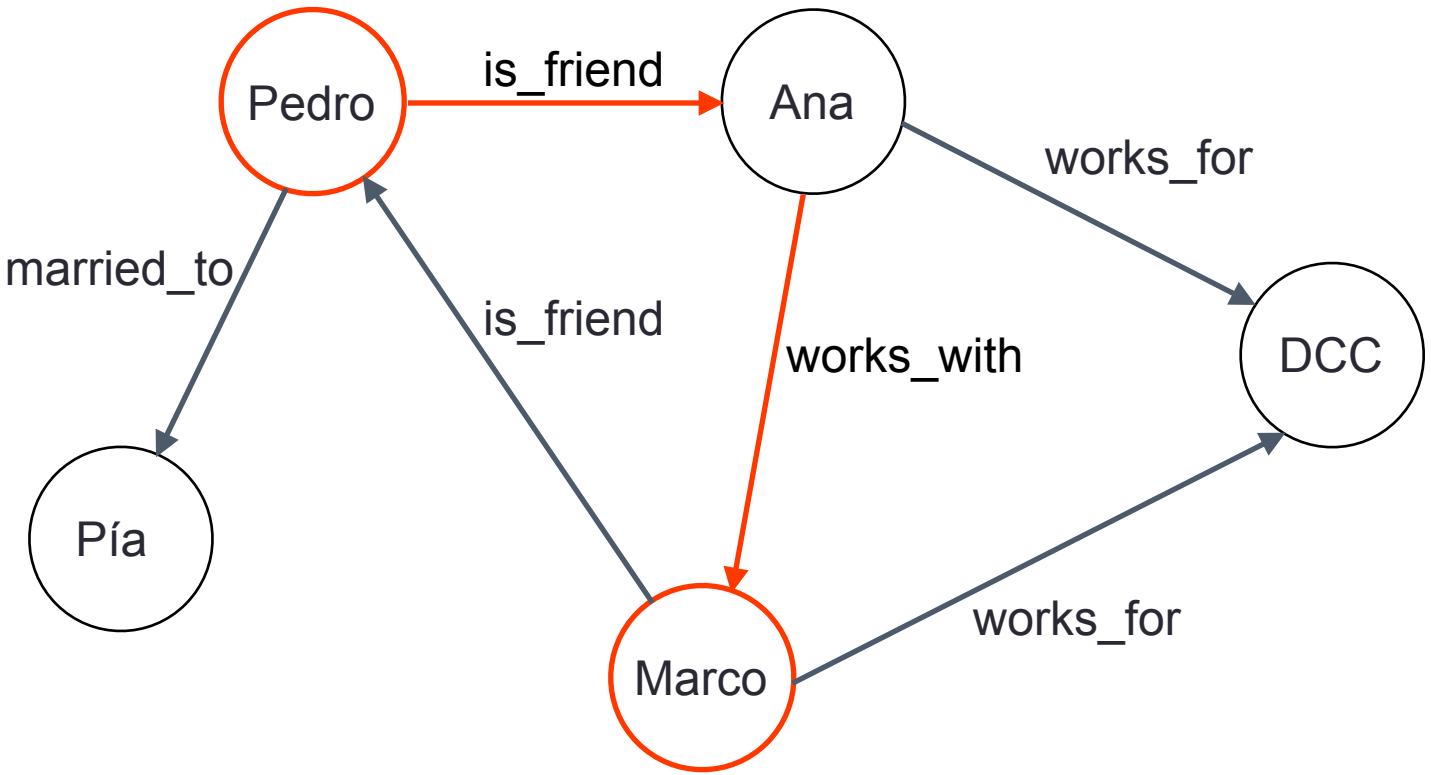
Esta charla

- Regular Path Queries (RPQs)
- Extensiones: Conjunciones y proyecciones (CRPQs)
- Patrones de grafos, aplicaciones

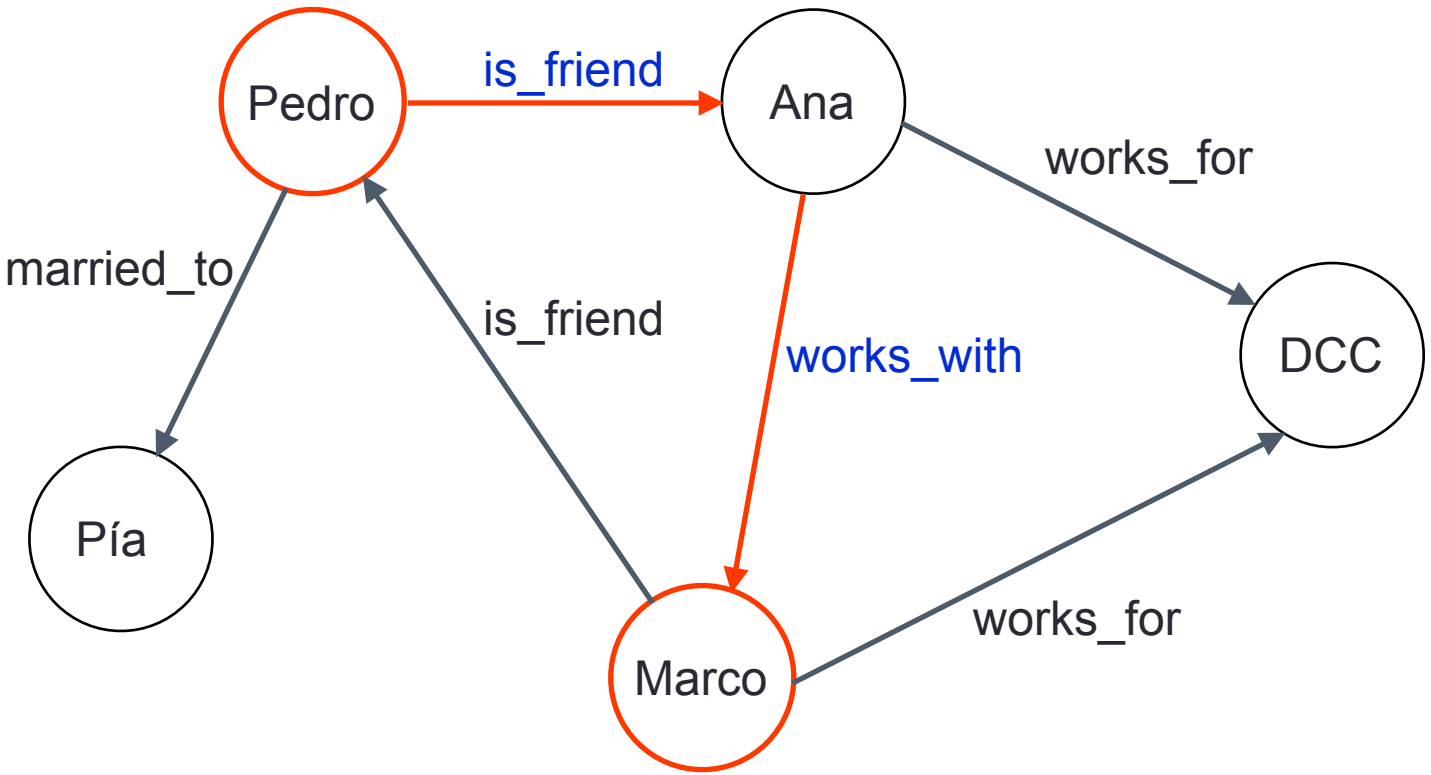
Esta charla

- Regular Path Queries (RPQs)
- Extensiones: Conjunciones y proyecciones (CRPQs)
- Patrones de grafos, aplicaciones

Un poco de notación…

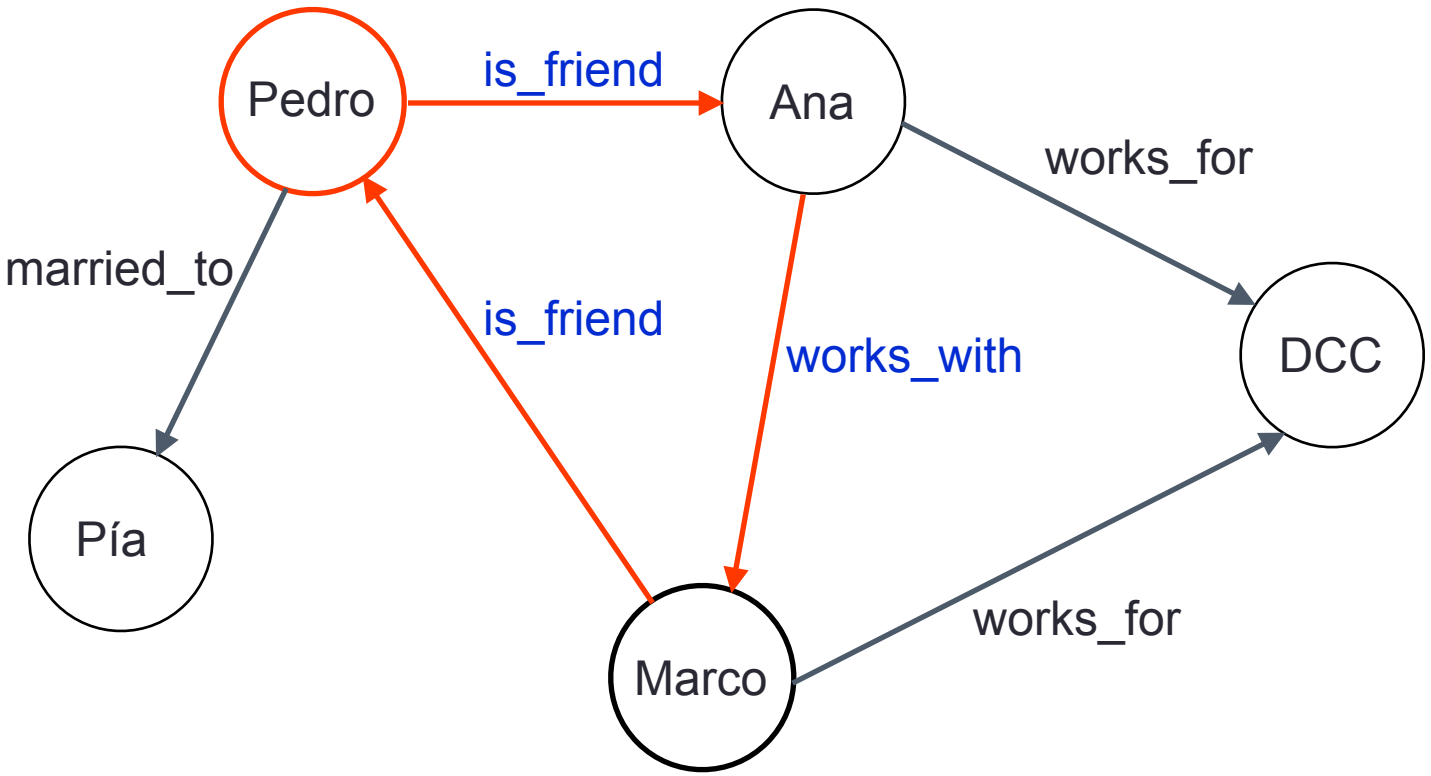


- Camino entre Pedro y Marco (pasa por Ana)

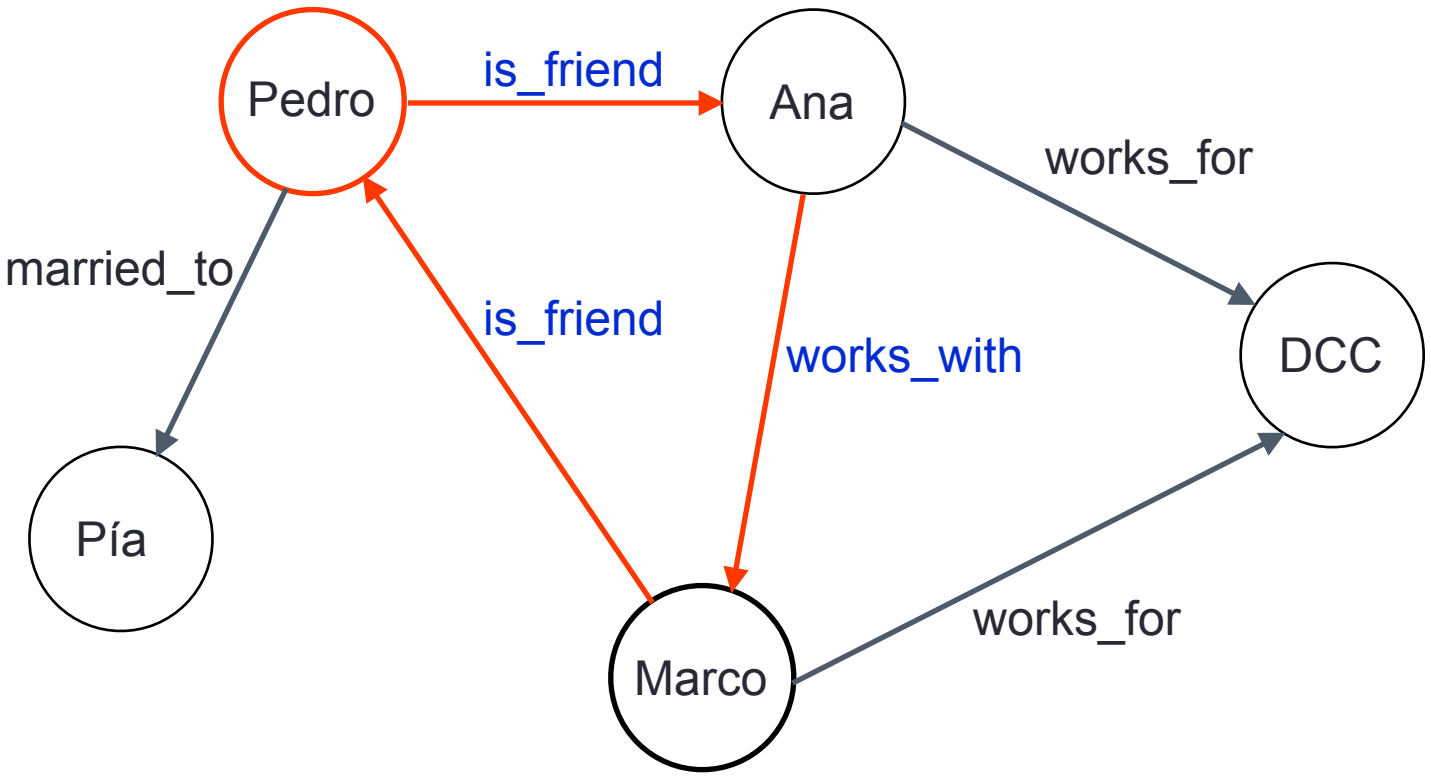


- Camino entre Pedro y Marco (pasa por Ana)
- La etiqueta del camino es

  is_friend . works_with

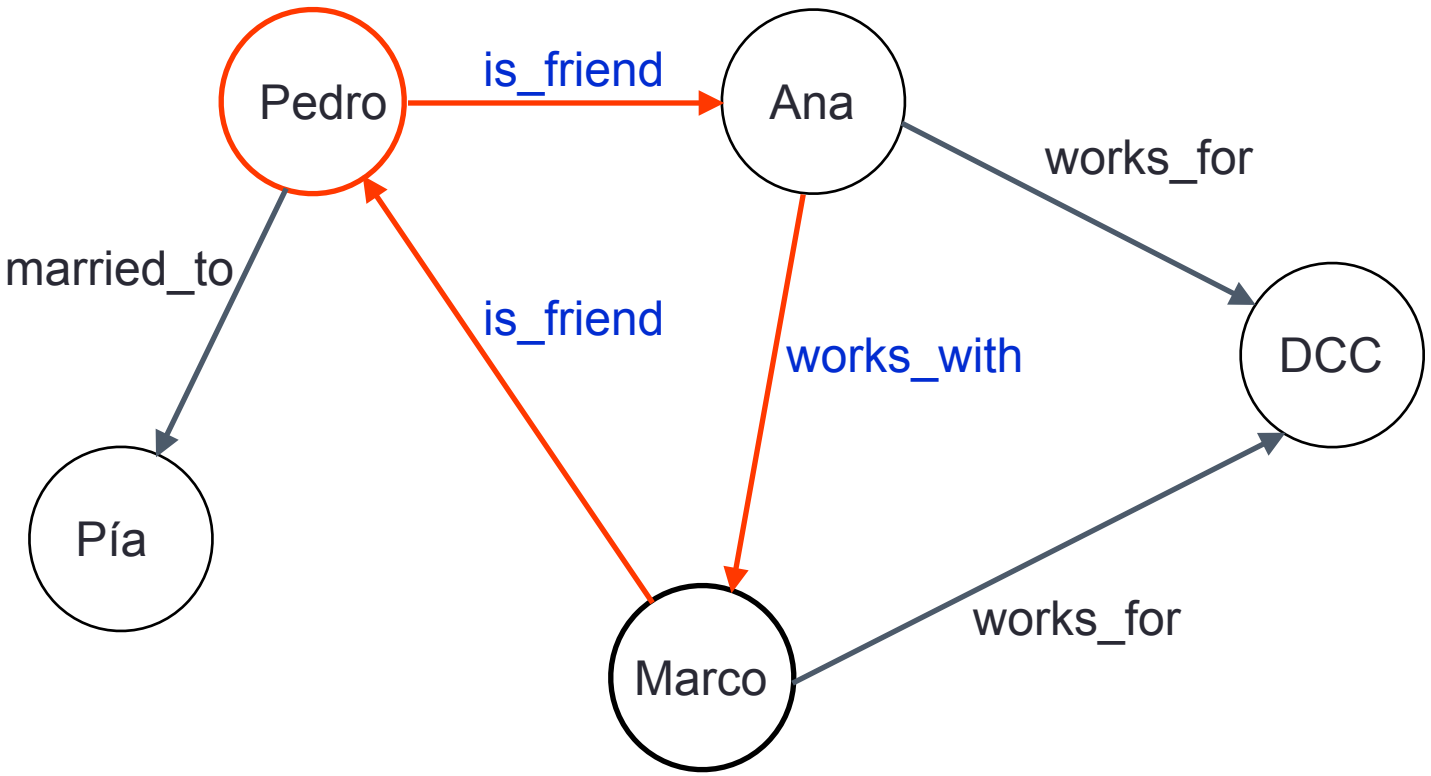


- Infinitos camino entre Pedro y Pedro (ciclo)

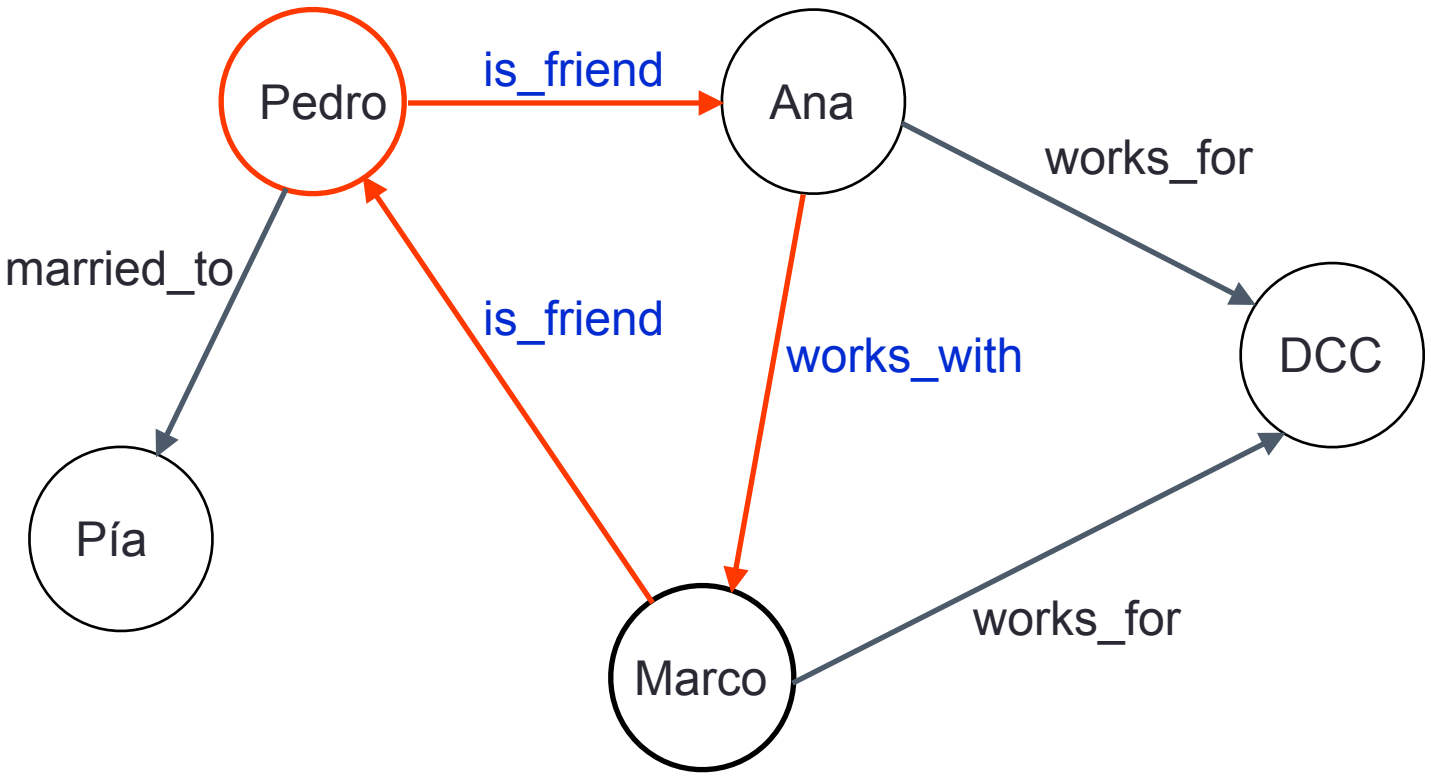


- Infinitos camino entre Pedro y Pedro (ciclo)
- Algunos de estos caminos:

is_friend . works_with . is_friend
is_friend . works_with . is_friend . is_friend . works_with . is_friend



- Las Regular Path Queries (RPQs) extraen información de caminos dados por lenguajes regulares



- Las Regular Path Queries (RPQs) extraen información de caminos dados por lenguajes regulares

- Entre Pedro y Pedro los caminos están dados por

  (is_friend . works_with . is_friend)*

Expresiones Regulares

Sea Σ un conjunto de etiquetas de aristas.
Definimos el conjunto de Expresiones Regulares sobre Σ:

- ε es una ER
- Toda etiqueta *a* en Σ es una ER
- Si r_1 y r_2 son ERs, entonces:
  - $r_1 \,.\, r_2$ es una ER
  - $r_1 + r_2$ es una ER
  - r_1^* es una ER

Expresiones Regulares: Semántica

Las expresiones regulares definen conjuntos de secuencias de etiquetas (palabras)

- *a* define el conjunto $\{a\}$
- *a* . *b* define el conjunto $\{ab\}$

Expresiones Regulares: Semántica

Las expresiones regulares definen conjuntos de secuencias de etiquetas (palabras)

- $r_1 + r_2$ define la union de las secuencias de r_1 y r_2
- $a + a.b$ es $\{a, ab\}$
- $(a + b).(c + d)$ es $\{ac, ad, bc, bd\}$

Expresiones Regulares: Semántica

Las expresiones regulares definen conjuntos de secuencias de etiquetas (palabras)

- r^* define:

  $$\varepsilon \cup r \cup r.r \cup r.r.r \cup \ldots$$

- a^* define $\{\varepsilon, a, aa, aaa, aaaa \ldots\}$
- $(a + b)^*$ son todas las palabras que puedo formar con a y b

RPQs

Seleccionan pares de nodos que estén conectados por un camino conforme a la expresión regular

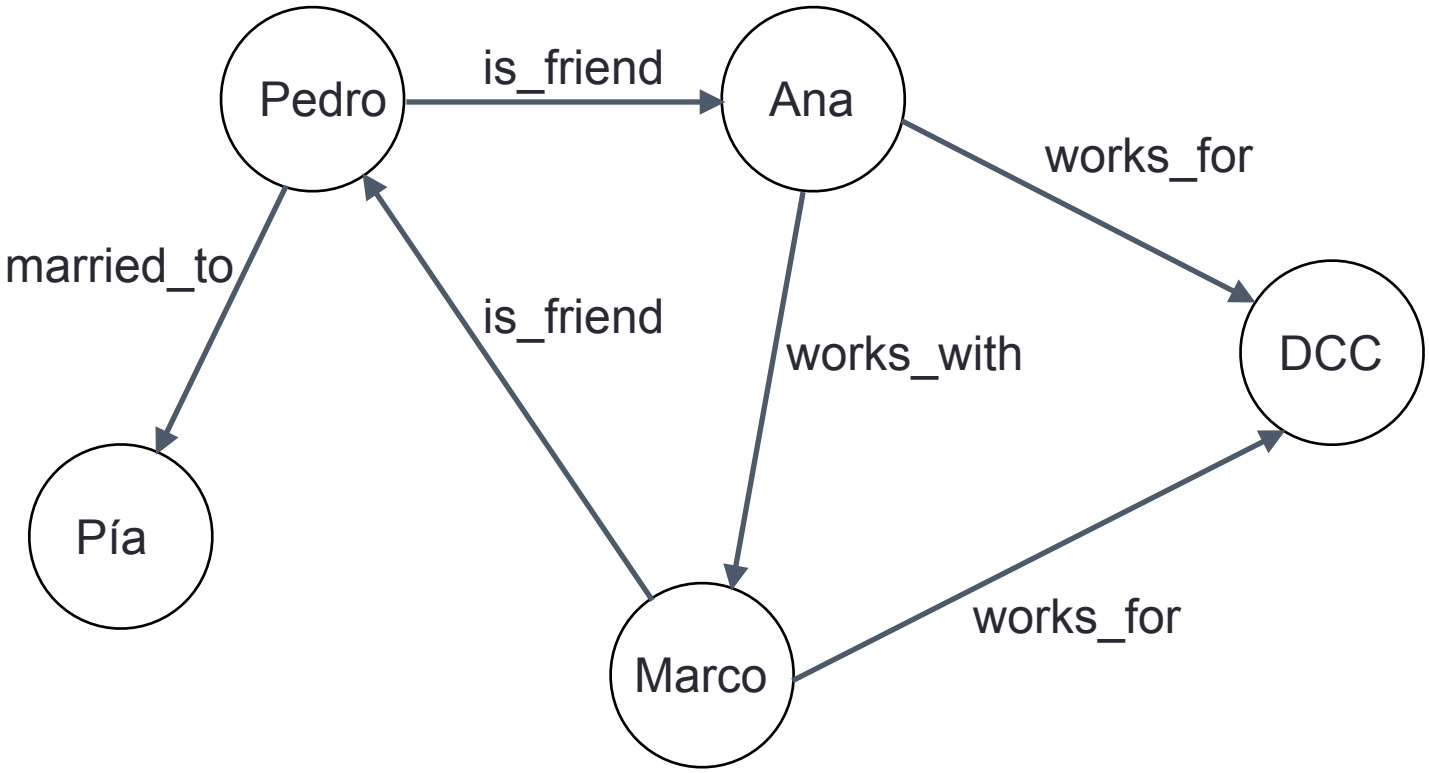
Formalmente, una RPQ sobre Σ es un par

(x,r,y)

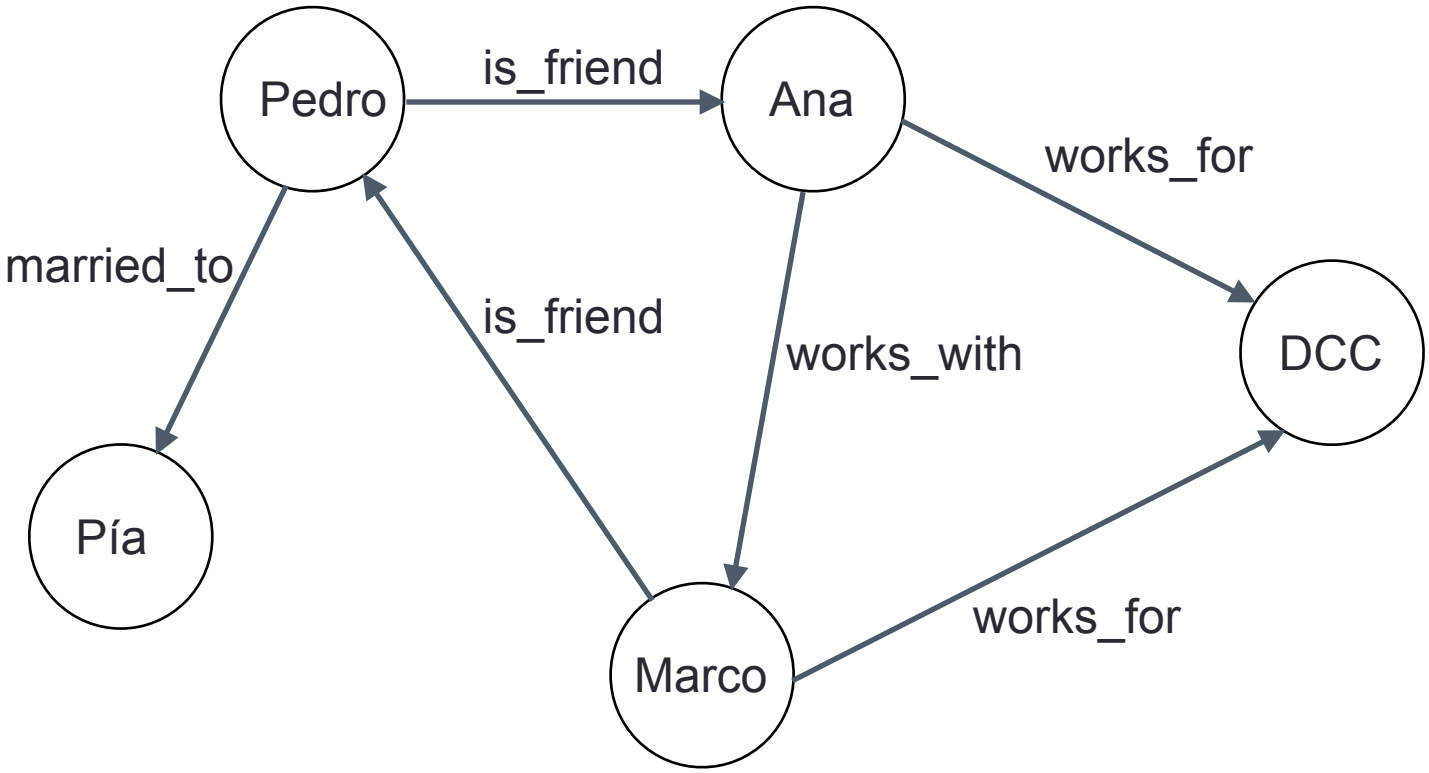
Donde r es una expresión regular sobre Σ

RPQs: Semántica

La RPQ (x,r,y) selecciona en un grafo todos los pares de nodos (x,y) tales que existe un camino etiquetado por r en el grafo

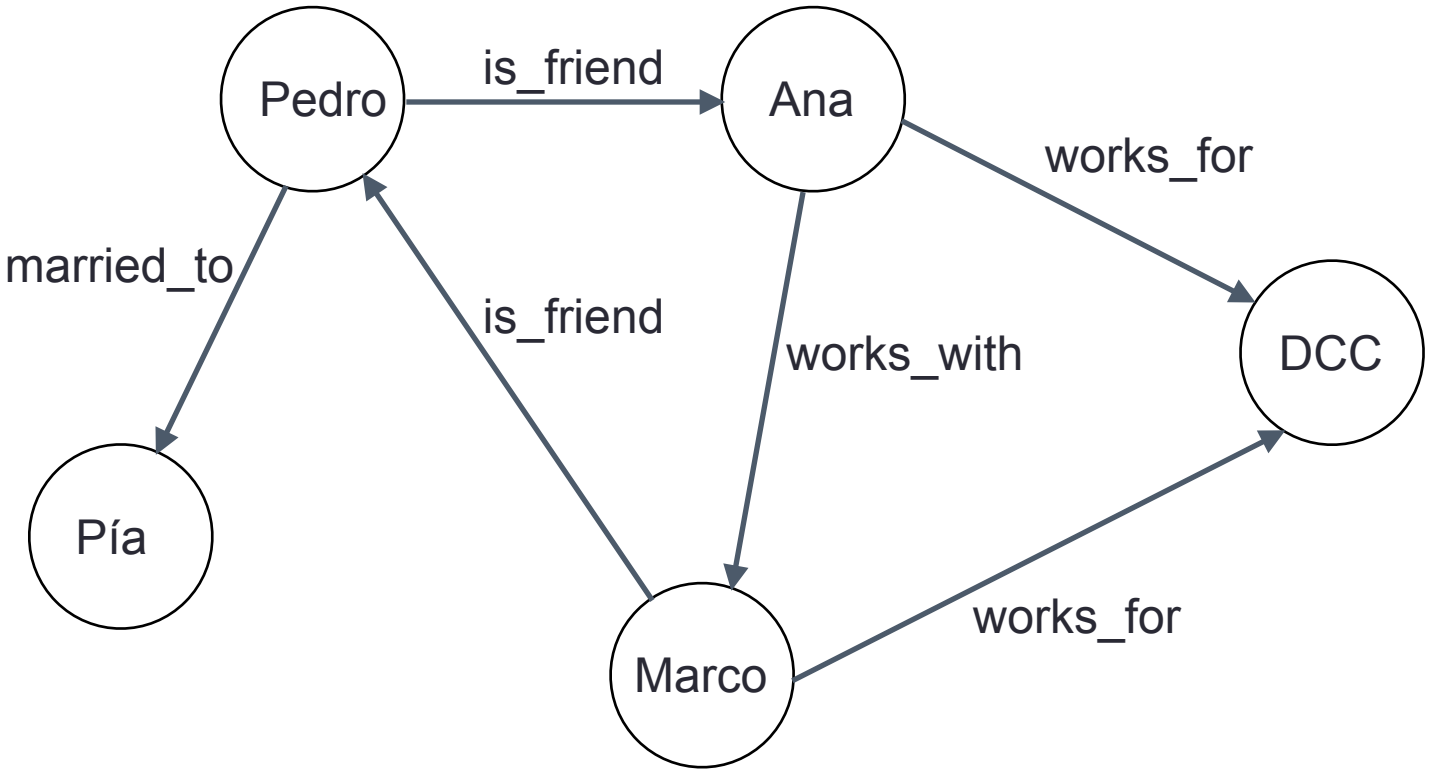


(x, is_friend . works_with, y) selecciona:

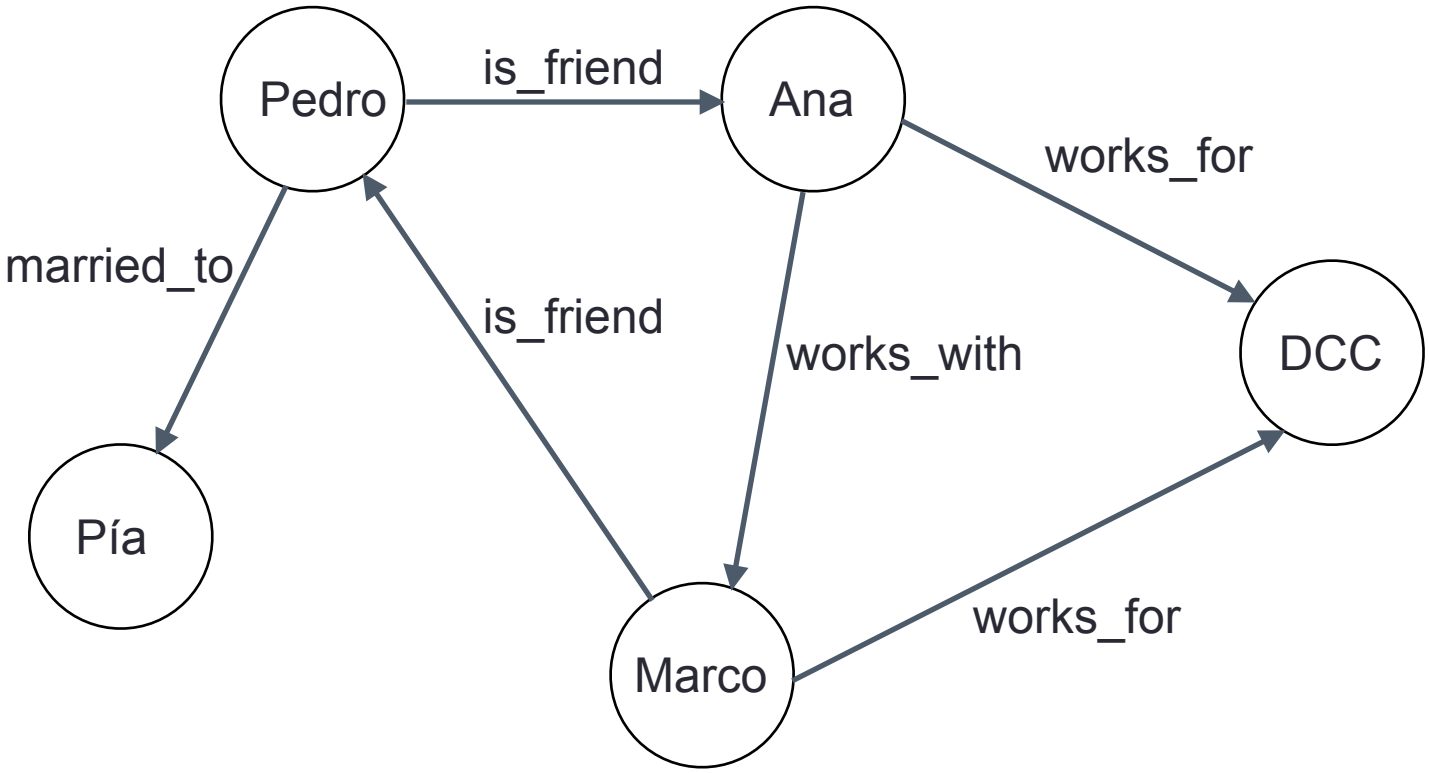


(x, is_friend . works_with, y) selecciona:

(Pedro, Marco)

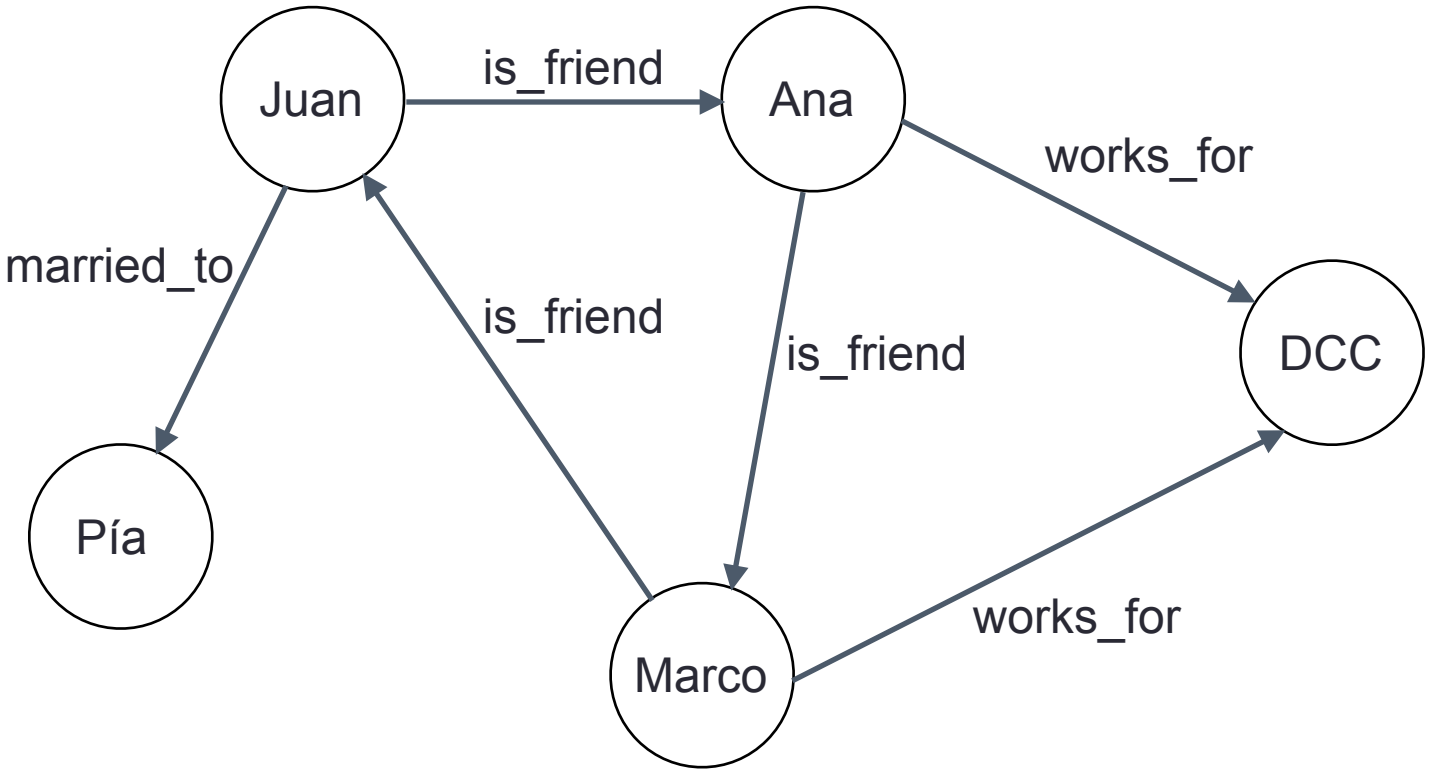


(x, is_friend . (works_with + married_to), y) selecciona:

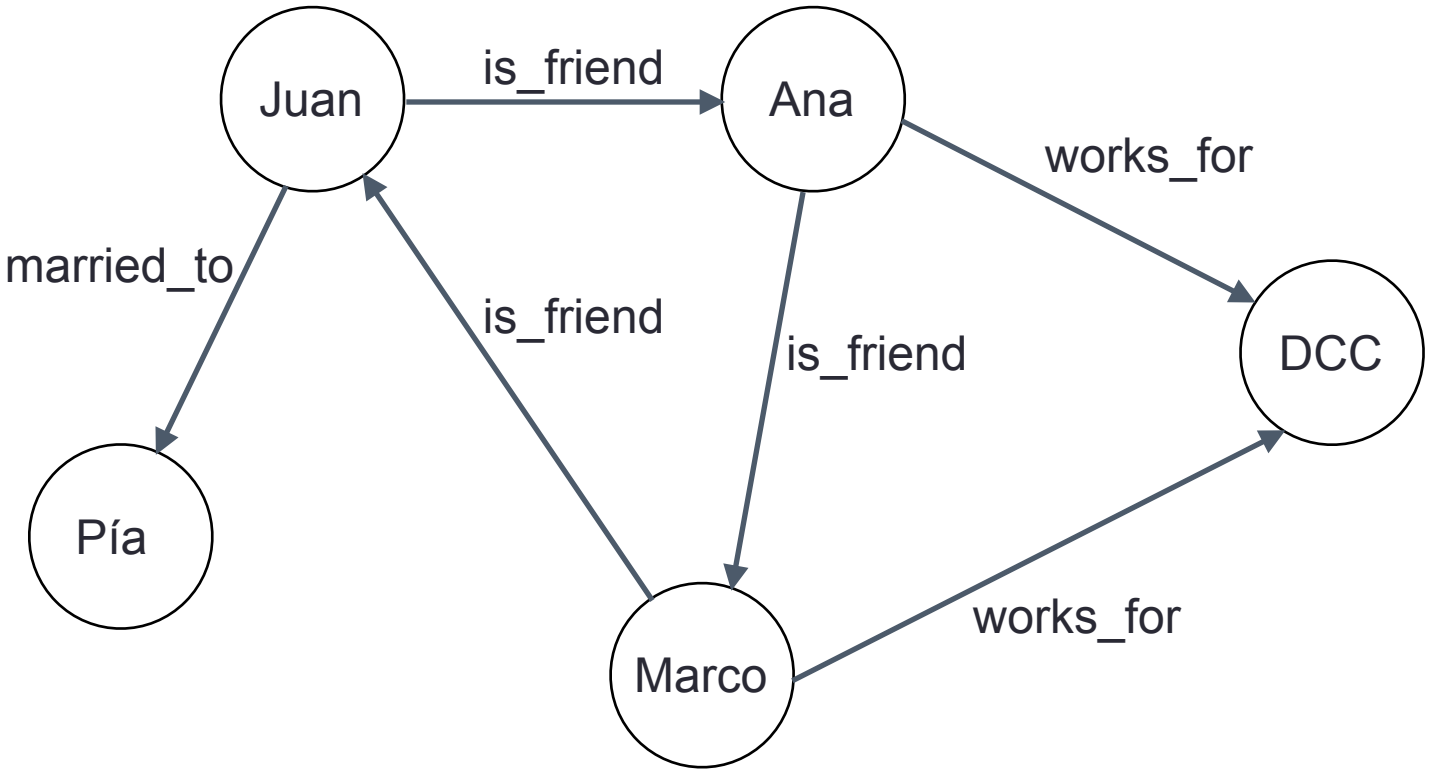


(x, is_friend . (works_with + married_to), y) selecciona:

(Pedro, Marco)
(Marco,Pia)

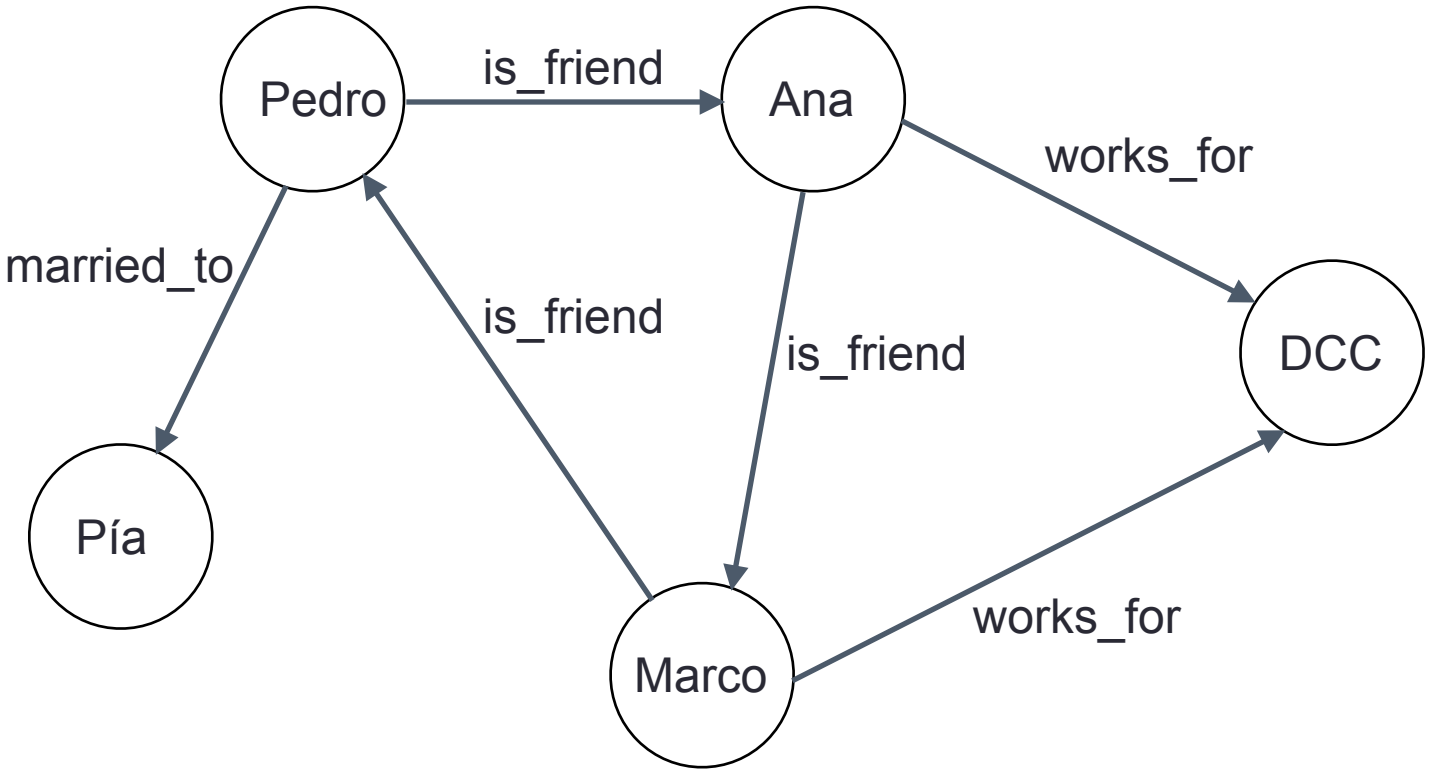


(x, is_friend*, y) selecciona:

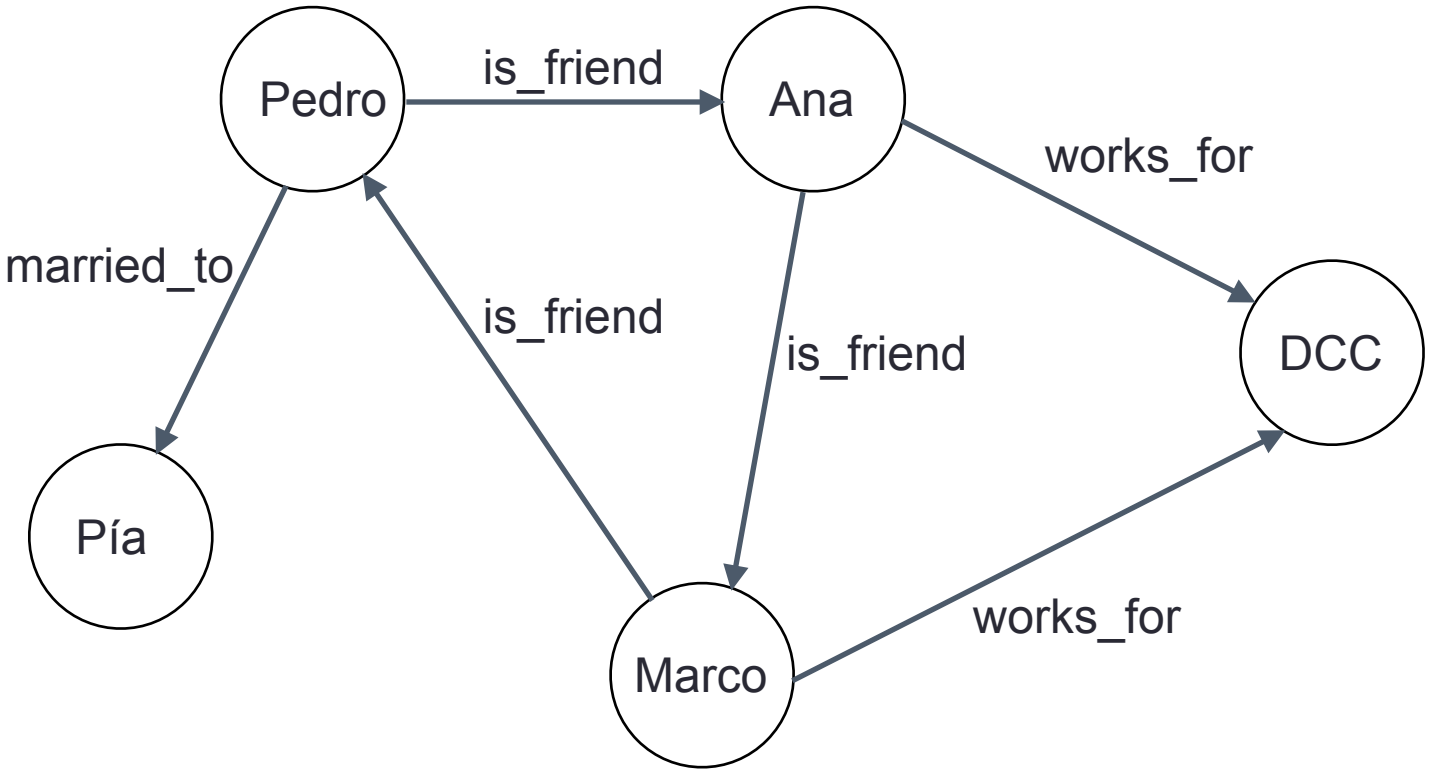


(x, is_friend*, y) selecciona:

(Pedro, Pedro) (Pedro, Ana) (Pedro, Marco)
(Ana, Ana) (Ana, Pedro) (Ana, Marco)
(Marco, Marco) (Marco, Pedro) (Marco, Ana)



(x, is_friend*.works_for, y) selecciona:



(x, is_friend*.works_for, y) selecciona:

(Pedro, DCC) (Ana,DCC) (Marco,DCC)

Las RPQs son hoy implementadas en muchos sistemas de Bases de Datos de grafos

Las RPQs son hoy implementadas en muchos sistemas de Bases de Datos de grafos

El estudio de las RPQs está fuertemente relacionado con teoría de automatas

Problema: Computar la respuesta a las RPQs

Dado un grafo *G* y una RPQ *r*,
ambos con etiquetas en Σ.

Cómo calculo los pares de nodos seleccionados por *r*?

¡No podemos hacer fuerza bruta!

- Algoritmo: chequear todos los caminos entre los nodos, ver si alguno está representado por la expresión.

¡No podemos hacer fuerza bruta!

- Algoritmo: chequear todos los caminos entre los nodos, ver si alguno está representado por la expresión.

Si hay ciclos tengo infinitos caminos, no puedo chequear 1 a 1.

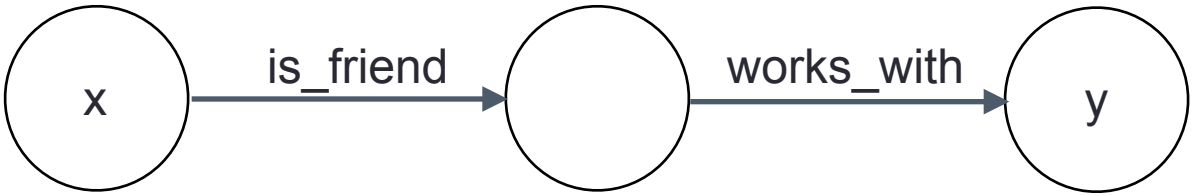
Teoría de autómatas al rescate

Teoría de autómatas al rescate

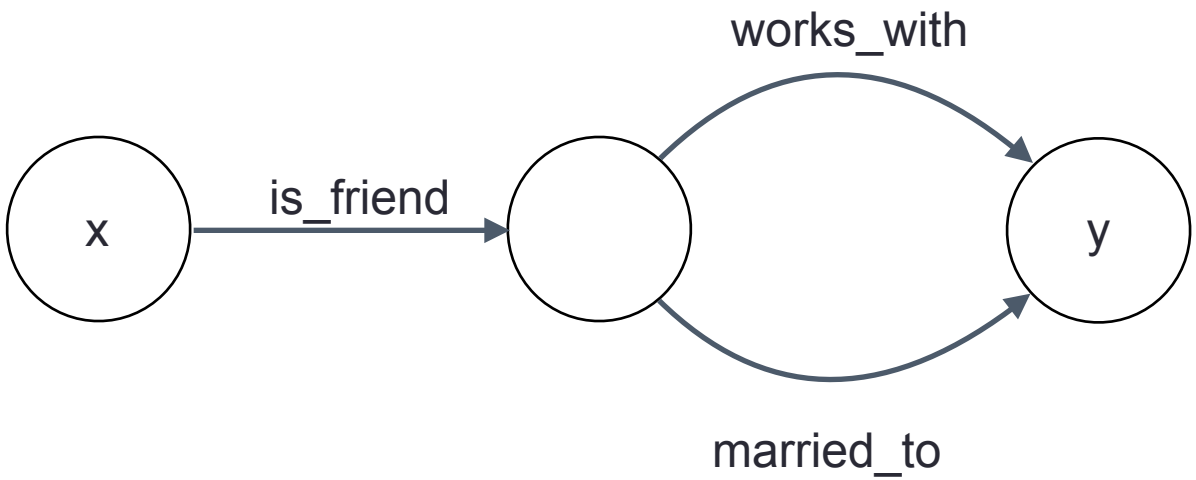
Idea:

Podemos representar también las RPQs como grafos

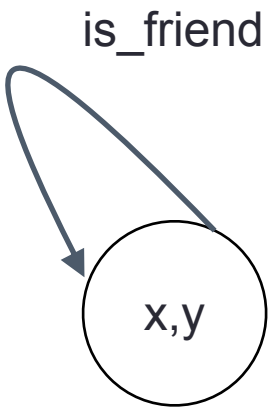
(x, is_friend . works_with, y)



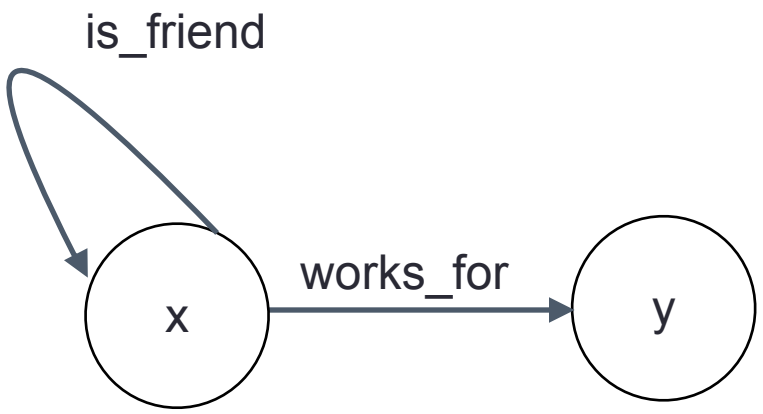
(x, is_friend . (works_with + married_to), y)



(x, is_friend*, y)



(x, is_friend*.works_for, y)



Teoría de autómatas al rescate

Formalmente:

Para cada RPQ (x,r,y) construimos grafo con nodos x e y tal que:

Teoría de autómatas al rescate

Formalmente:

Para cada RPQ (x,r,y) construimos grafo con nodos x e y tal que:

Los caminos entre x e y en el grafo
corresponden a las secuencias representadas por r

Teoría de autómatas al rescate

Formalmente:

Para cada RPQ (x,r,y) construimos grafo con nodos x e y tal que:

Los caminos entre x e y en el grafo
corresponden a las secuencias representadas por r

- En computación estos objetos se llaman autómatas

Teoría de autómatas al rescate

Algoritmo para evaluar una RPQ (x,r,y) sobre un grafo G:

1. Construyo grafo R con nodos x e y

Teoría de autómatas al rescate

Algoritmo para evaluar una RPQ (x,r,y) sobre un grafo G:

1. Construyo grafo R con nodos x e y
2. Tomo el producto cruz de G con R.

Teoría de autómatas al rescate

Algoritmo para evaluar una RPQ (x,r,y) sobre un grafo G:

1. Construyo grafo R con nodos x e y
2. Tomo el producto cruz de G con R.
3. Veo si en G x R puedo alcanzar un nodo (u,y) desde un nodo (u,x), para u y v nodos de G

Teoría de autómatas al rescate

Algoritmo para evaluar una RPQ (x,r,y) sobre un grafo G:

1. Construyo grafo R con nodos x e y
2. Tomo el producto cruz de G con R.
3. Veo si en G x R puedo alcanzar un nodo (u,y) desde un nodo (u,x), para u y v nodos de G

Para el paso 3 usamos algoritmo de Dijkstra u otro similar

1. Construyo grafo R con nodos x e y

2. Tomo el producto cruz de G con R.

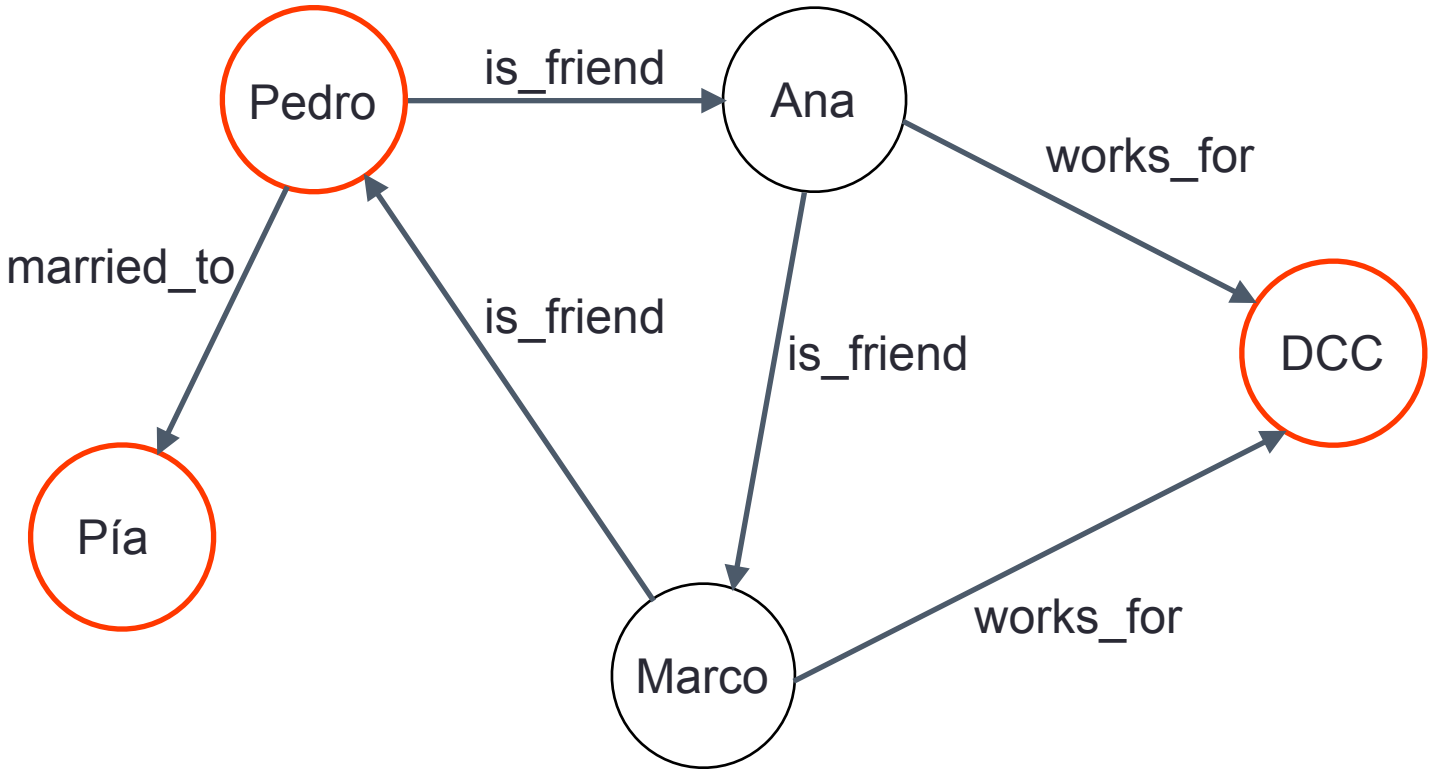
3. Veo si en G x R puedo alcanzar un nodo (u,y) desde un nodo (u,x), para u y v nodos de G

Computar las respuestas a RPQs toma tiempo lineal con respecto al grafo

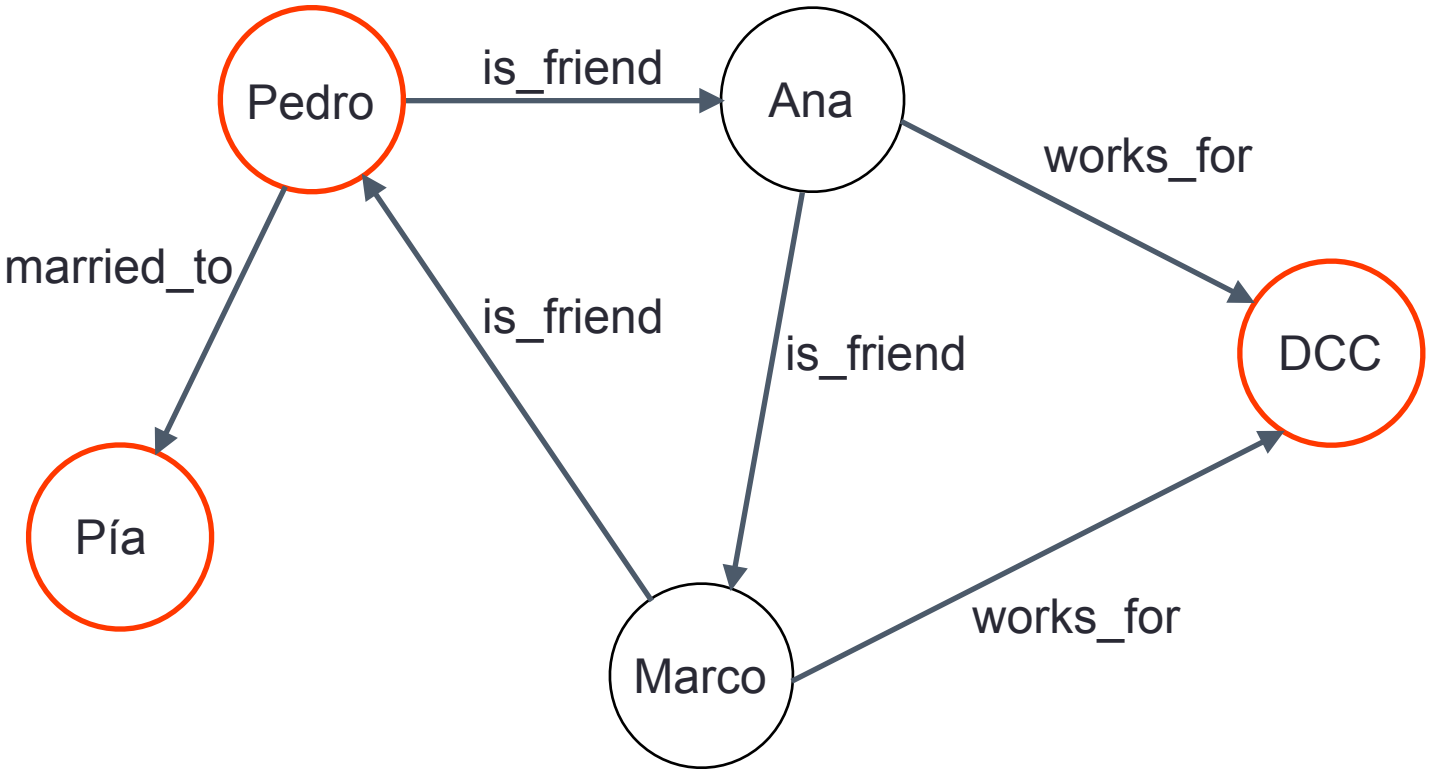
Esta charla

- Consultas de caminos regulares
- Extensiones: Conjunciones y proyecciones (CRPQs)
- Patrones de grafos, aplicaciones

(x, is_friend*.works_for, y) AND (x,married_to z)

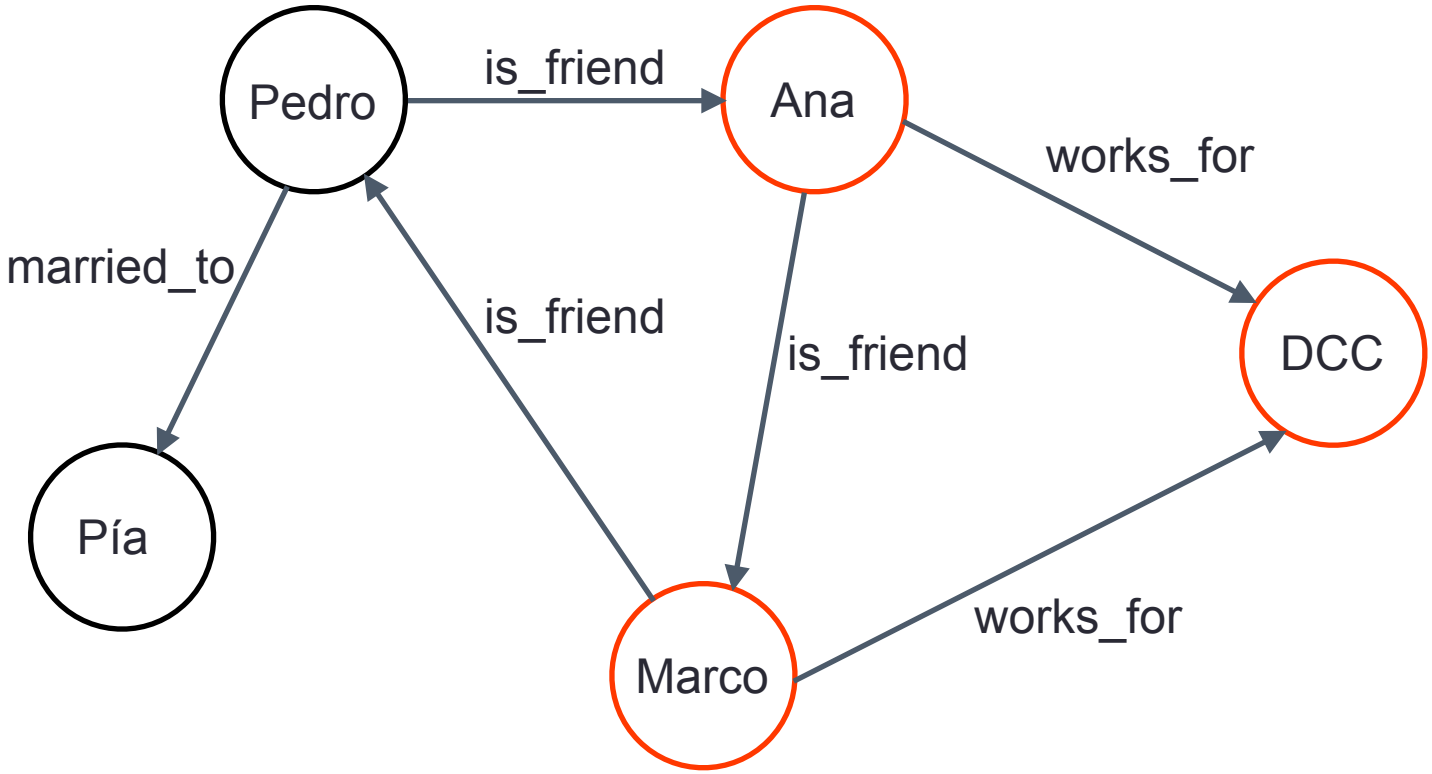


(x, is_friend*.works_for, y) AND (x,married_to z)

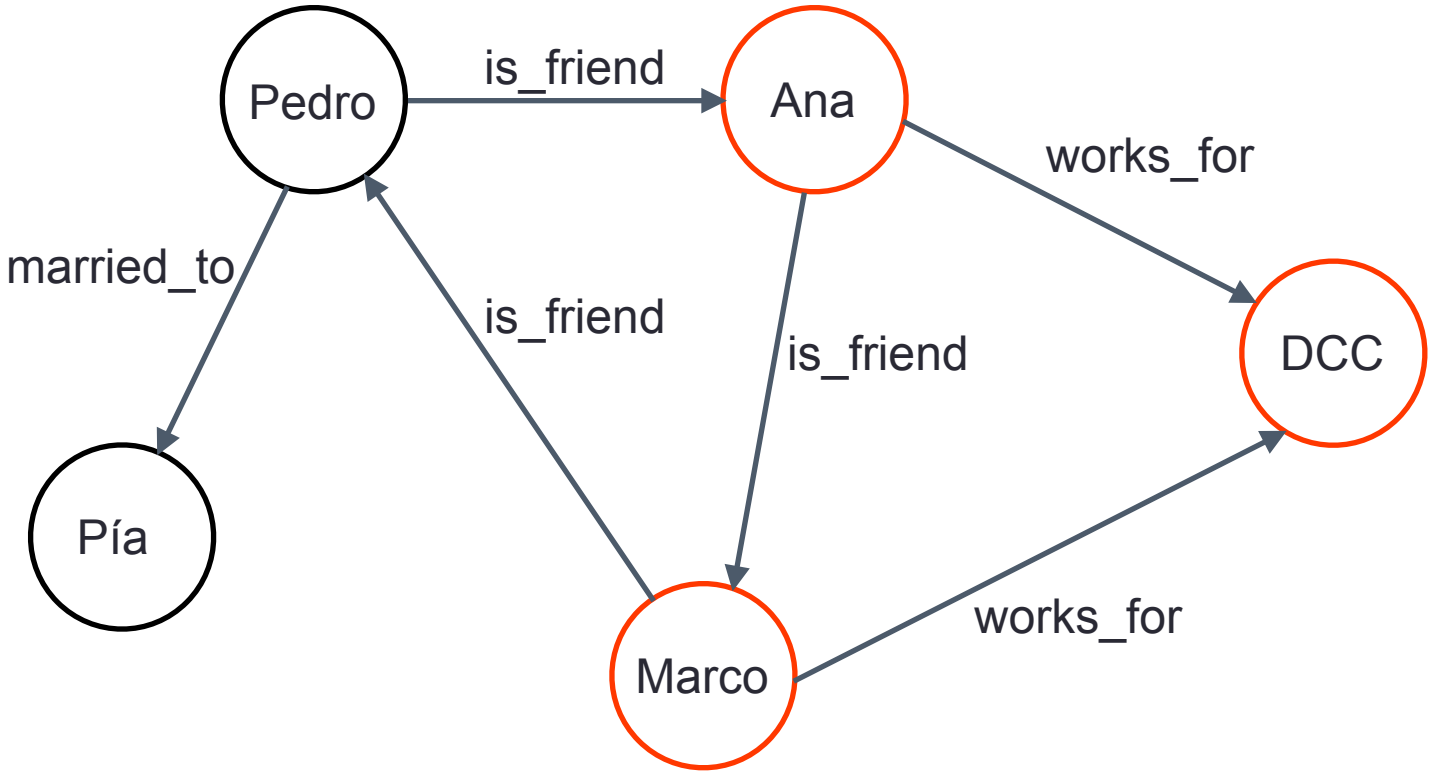


x	y	z
Pedro	DCC	Pía

(x, is_friend*, y) AND (x,works_for,z) AND (y,works_for,z)

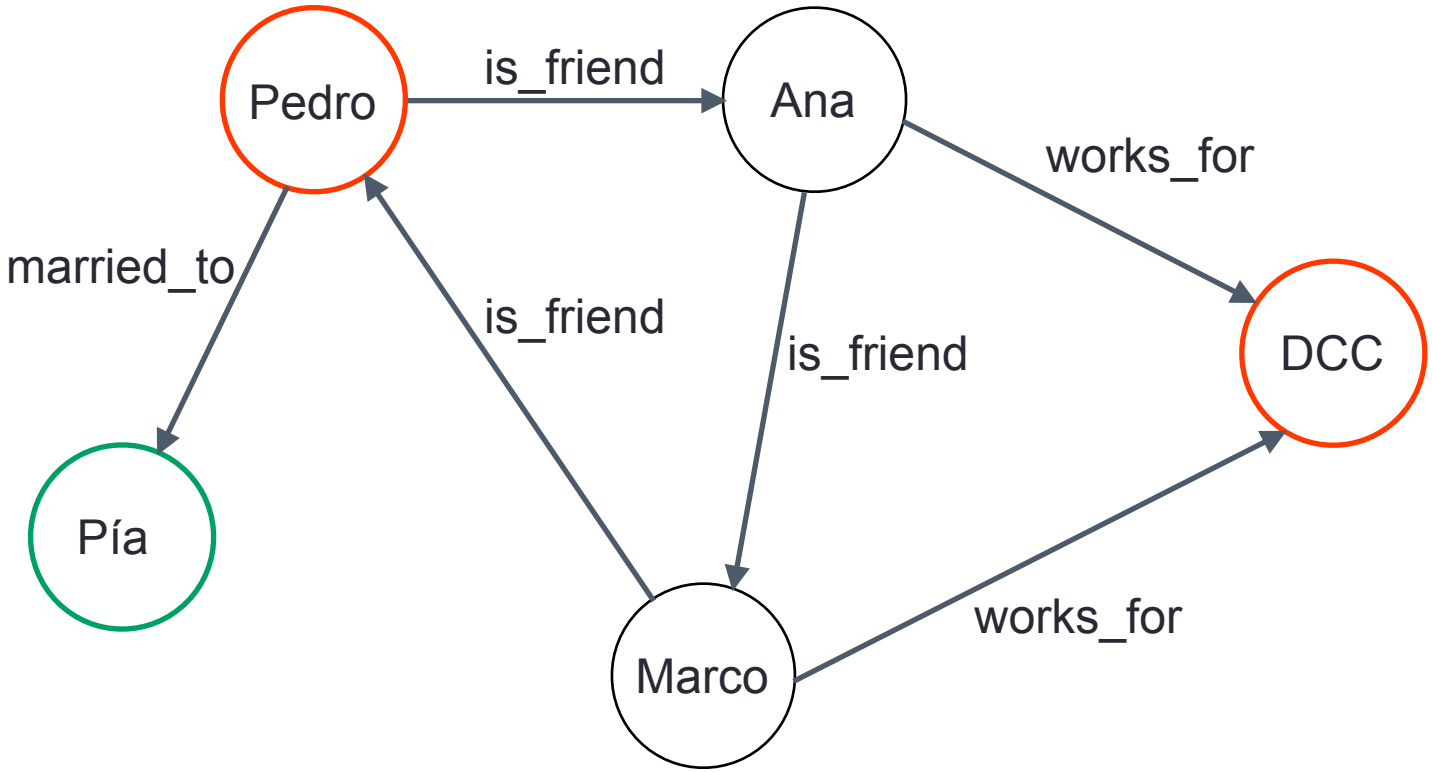


(x, is_friend*, y) AND (x,works_for,z) AND (y,works_for,z)



x	y	z
Ana	Marco	DCC
Marco	Ana	DCC

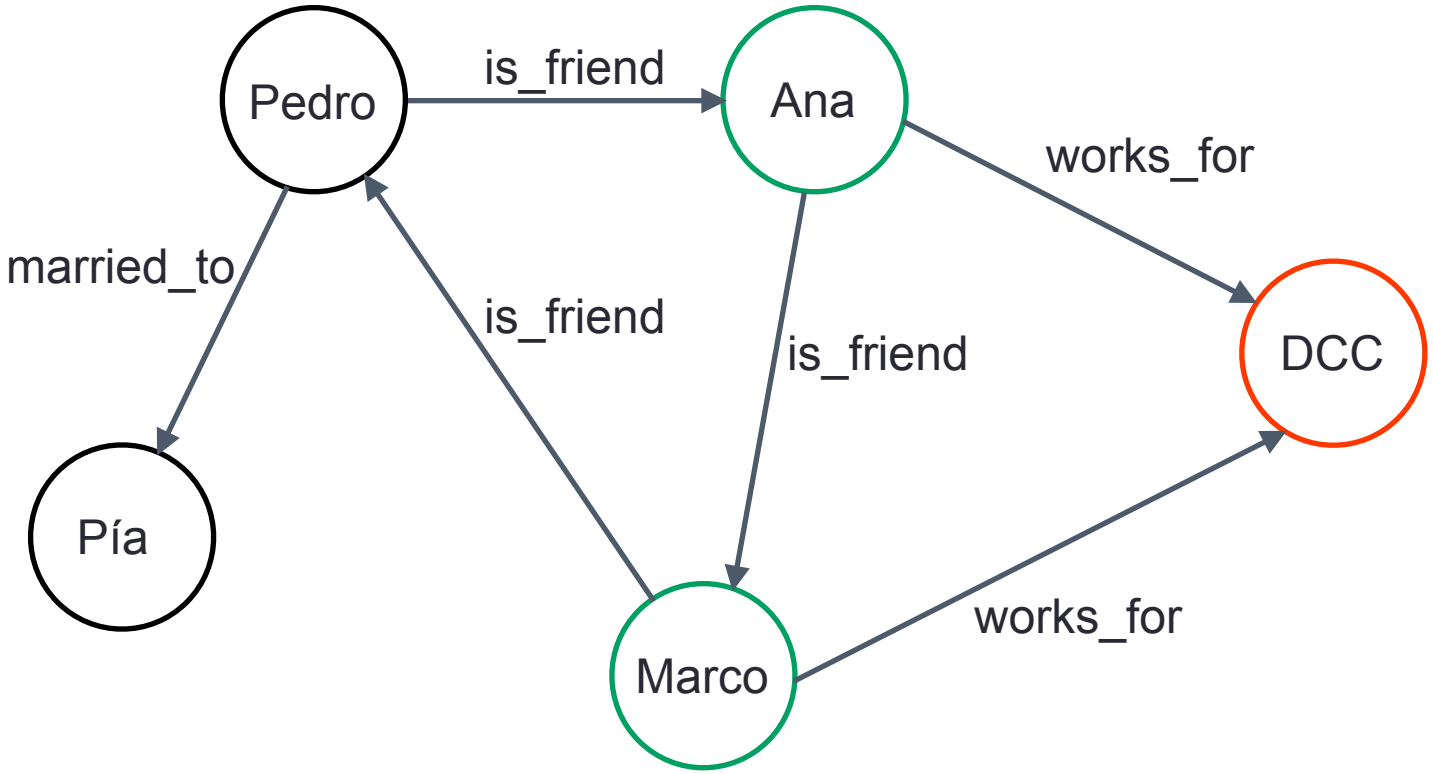
{ z | (x, is_friend*.works_for, y) AND (x,married_to z)}



z

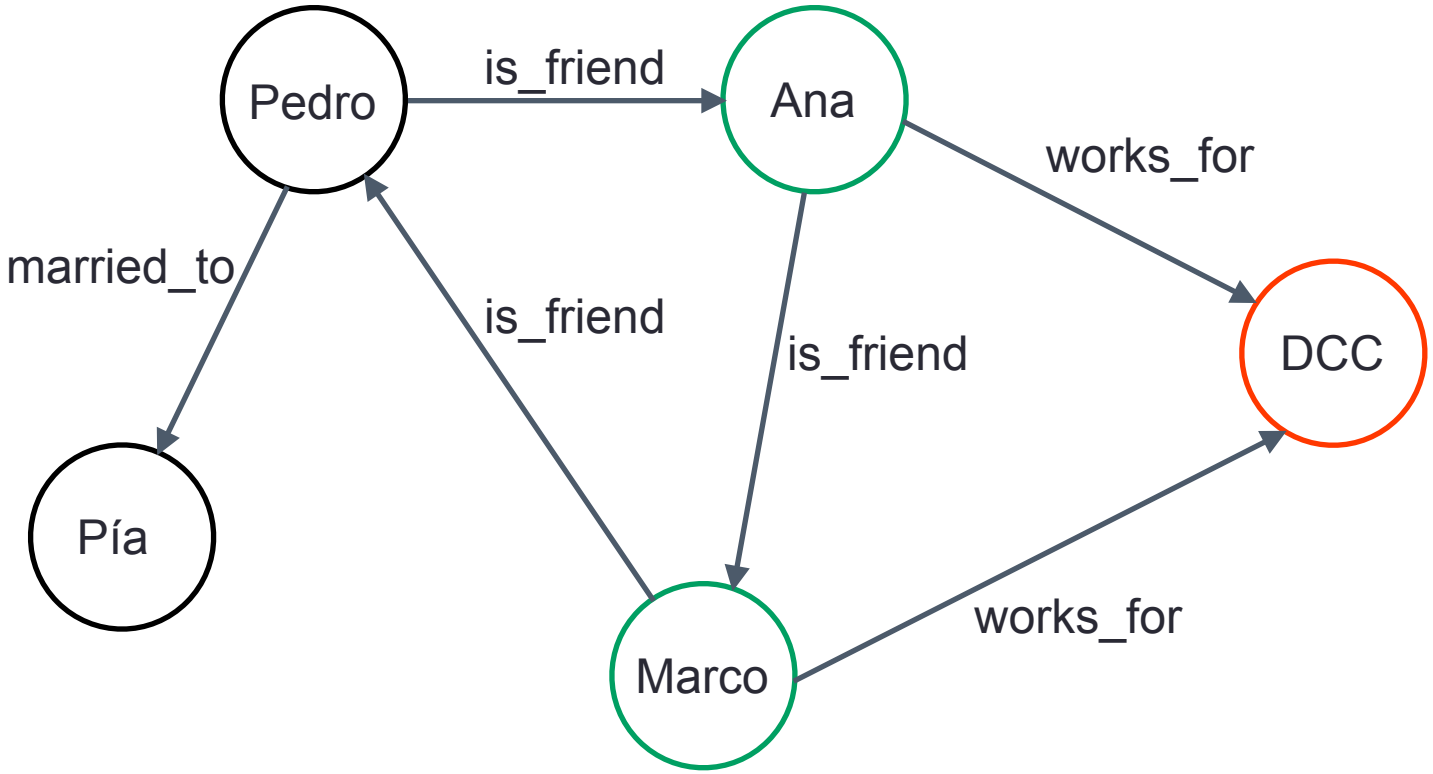
Pía

{x,y | (x, is_friend*, y) AND (x,works_for,z) AND (y,works_for,z)}



x	y
Ana	Marco
Marco	Ana

{x,y | (x, is_friend*, y) AND
(x,works_for,DCC) AND (y,works_for,DCC)}



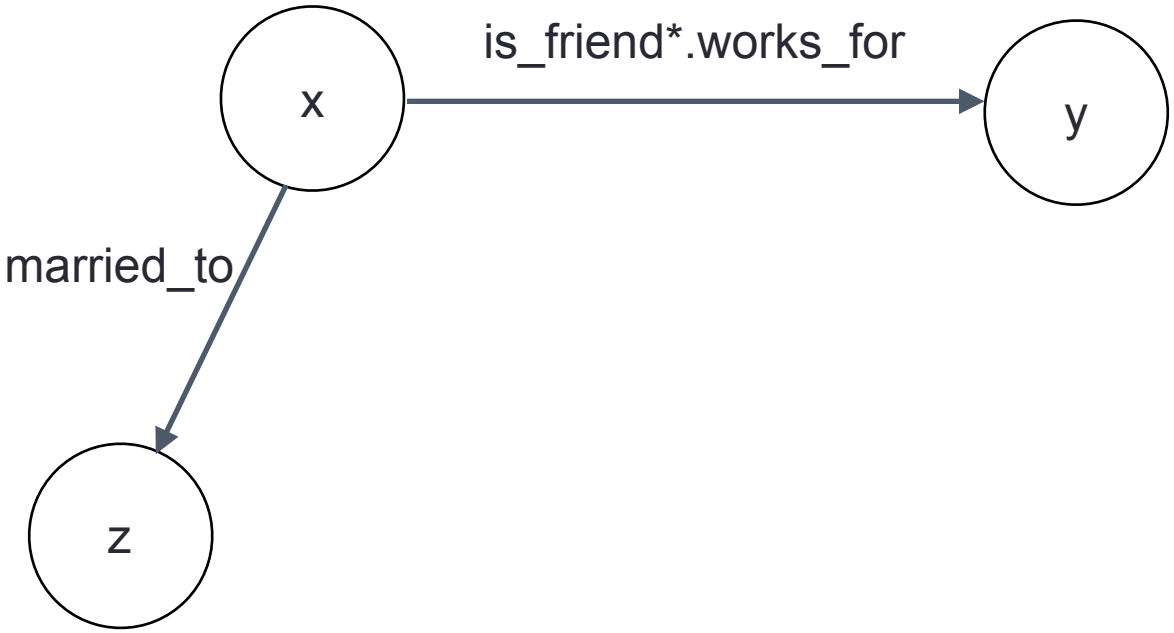
x	y
Ana	Marco
Marco	Ana

CRPQs son un buen lenguaje para grafos

- Contienen a las RPQs
- Contienen a los patrones
- Pueden ser expresadas como patrones más complejos

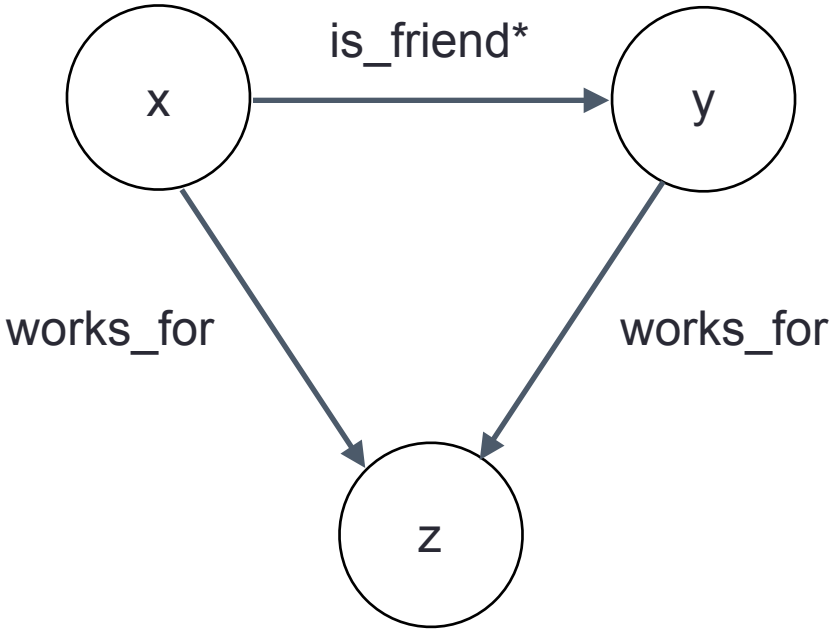
CRPQs como patrones más complejos

(x, is_friend*.works_for, y) AND (x,married_to z)



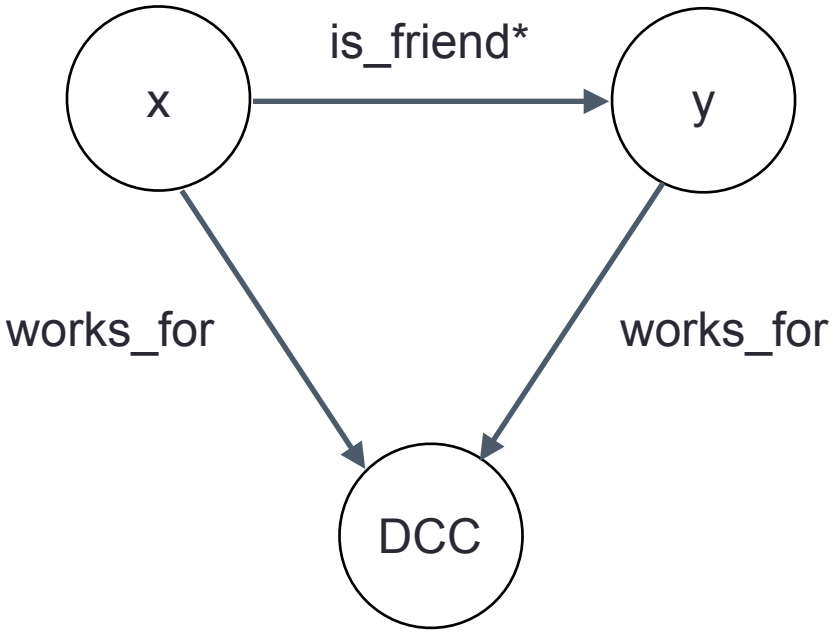
CRPQs como patrones más complejos

(x, is_friend*, y) AND (x,works_for,z) AND (y,works_for,z)



CRPQs como patrones más complejos

{x,y | (x, is_friend*, y) AND
(x,works_for,DCC) AND (y,works_for,DCC)}



CRPQs son un buen lenguaje para grafos

- Contienen a las RPQs
- Contienen a los patrones
- Pueden ser expresadas como patrones más complejos

¿Cómo puedo encontrar las respuestas a CRPQs?

Decidir si un conjunto de nodos es una respuesta a una CRPQ sobre un grafo es NP-completo

Decidir si un conjunto de nodos es una respuesta a una CRPQ sobre un grafo es NP-completo

Idea del algoritmo:

- combinar pattern matching con evaluación de RPQs

CRPQs son un buen lenguaje para grafos

Decidir si un conjunto de nodos es una respuesta a una CRPQ sobre un grafo es NP-completo

- Podemos agregar caminos a los patrones,
  sin costo de computación adicional

CRPQs son un buen lenguaje para grafos

Si consideramos la CRPQ como fija:

problema puede ser resuelto en tiempo polinomial

CRPQs son un buen lenguaje para grafos

Si consideramos la CRPQ como fija:

problema puede ser resuelto en tiempo polinomial

- Grafo es generalmente muy grande y consulta pequeña.
- Intuición: problema es robusto al tamaño del grafo

Esta charla

- Consultas de caminos regulares
- Extensiones: Conjunciones y proyecciones (CRPQs)
- Patrones de grafos, aplicaciones

Tenemos buenos
lenguajes de consulta para bases de datos de grafos

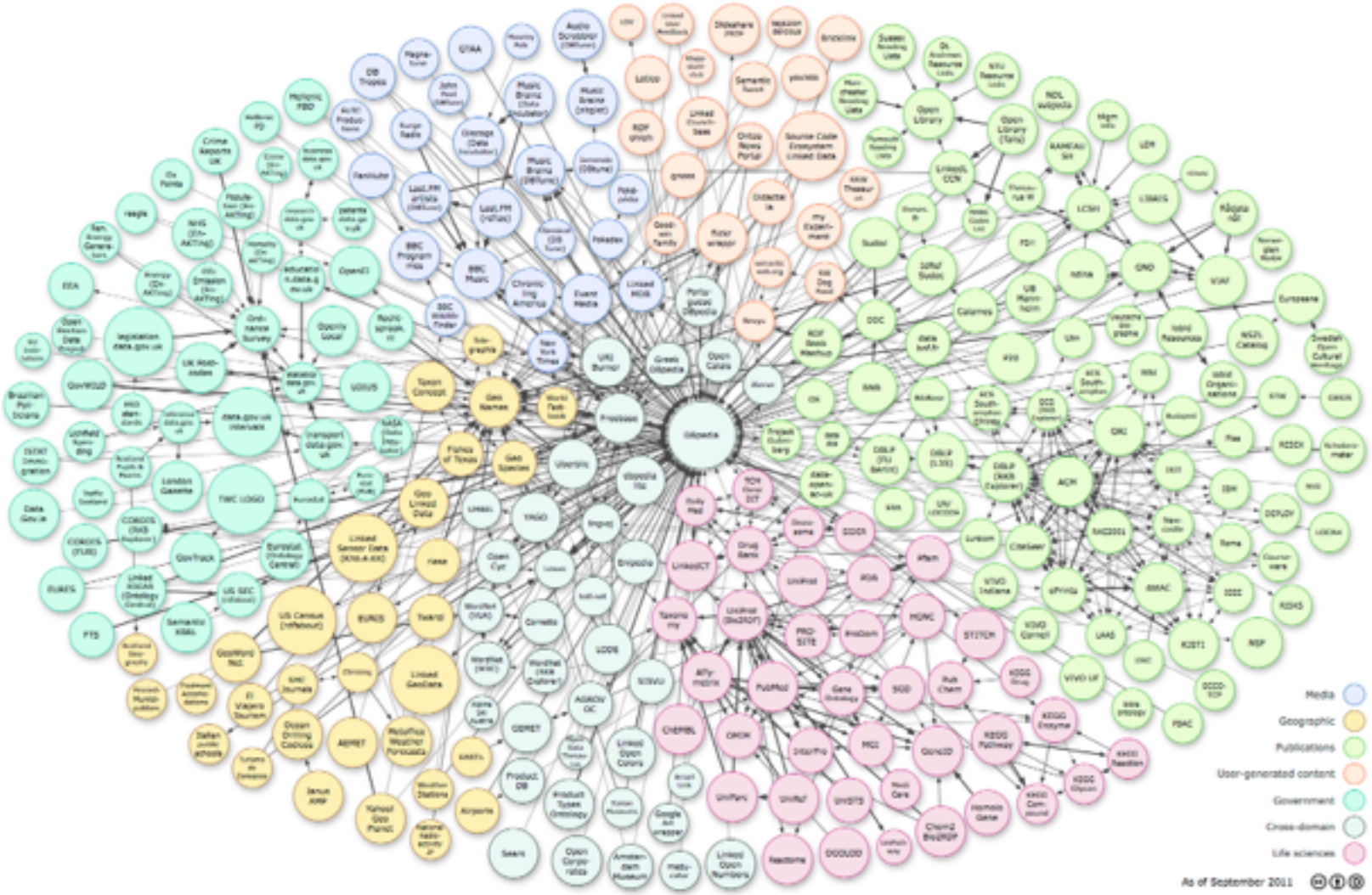
¿Qué cosas podemos hacer con esto?
(además de extraer información)

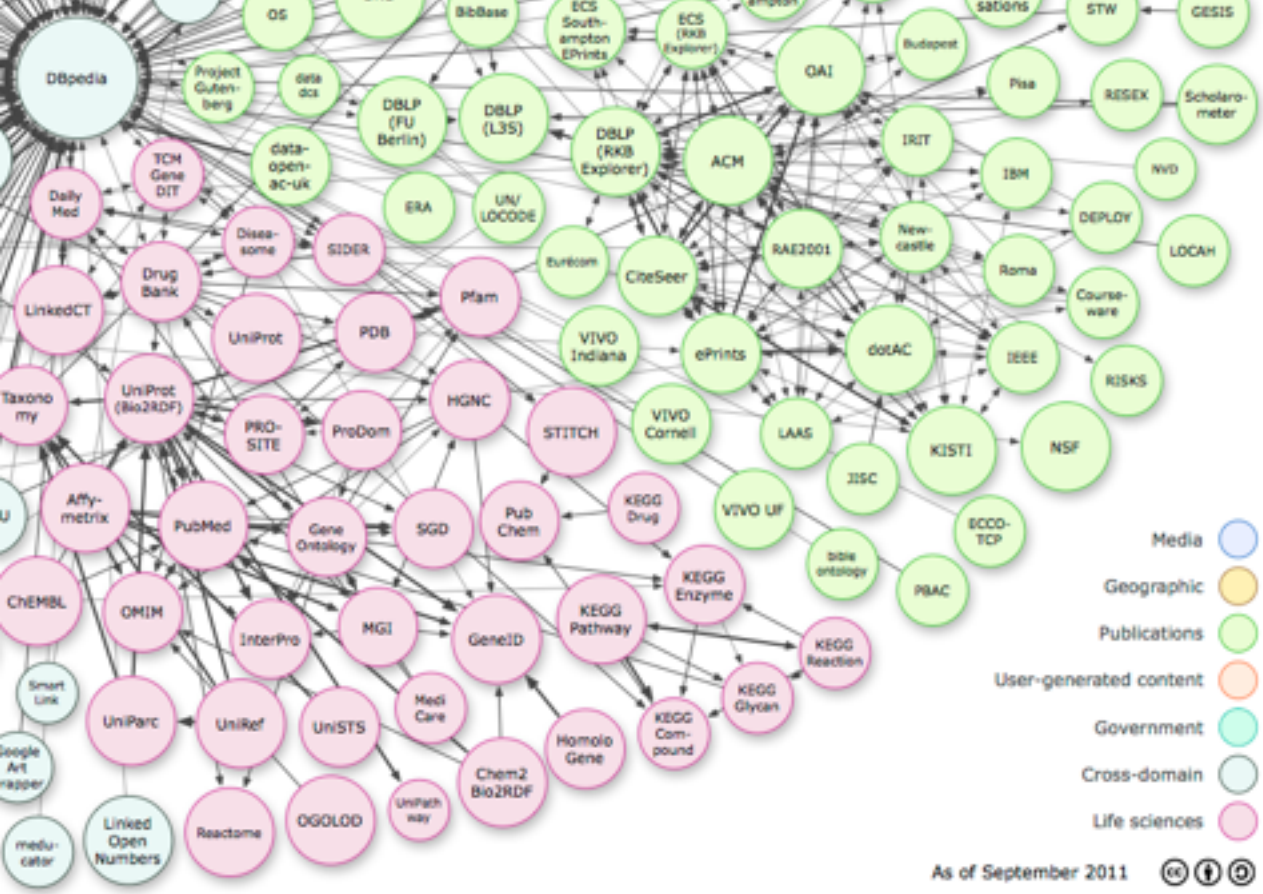
Tenemos buenos
lenguajes de consulta para bases de datos de grafos

¿Qué cosas podemos hacer con esto?
(además de extraer información)

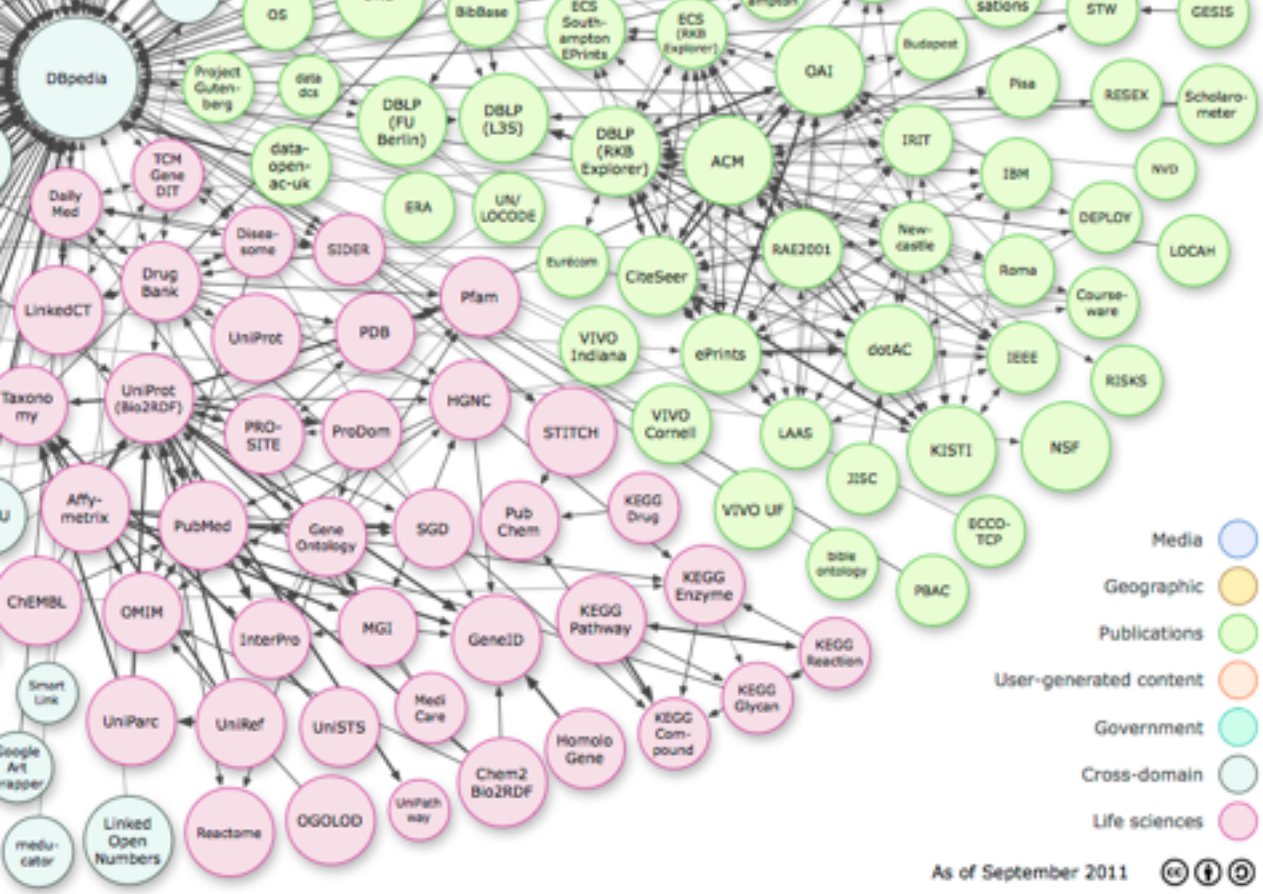
Usar estos lenguajes para representar
múltiples bases de datos de grafos

Linked Data


Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011



- Miles de bases de grafos conectadas entre sí
- Todas guardan información de forma differente



Necesitamos una forma de resumir e integrar grafos

Idea:

Resumamos un grafo usando un patrón complejo

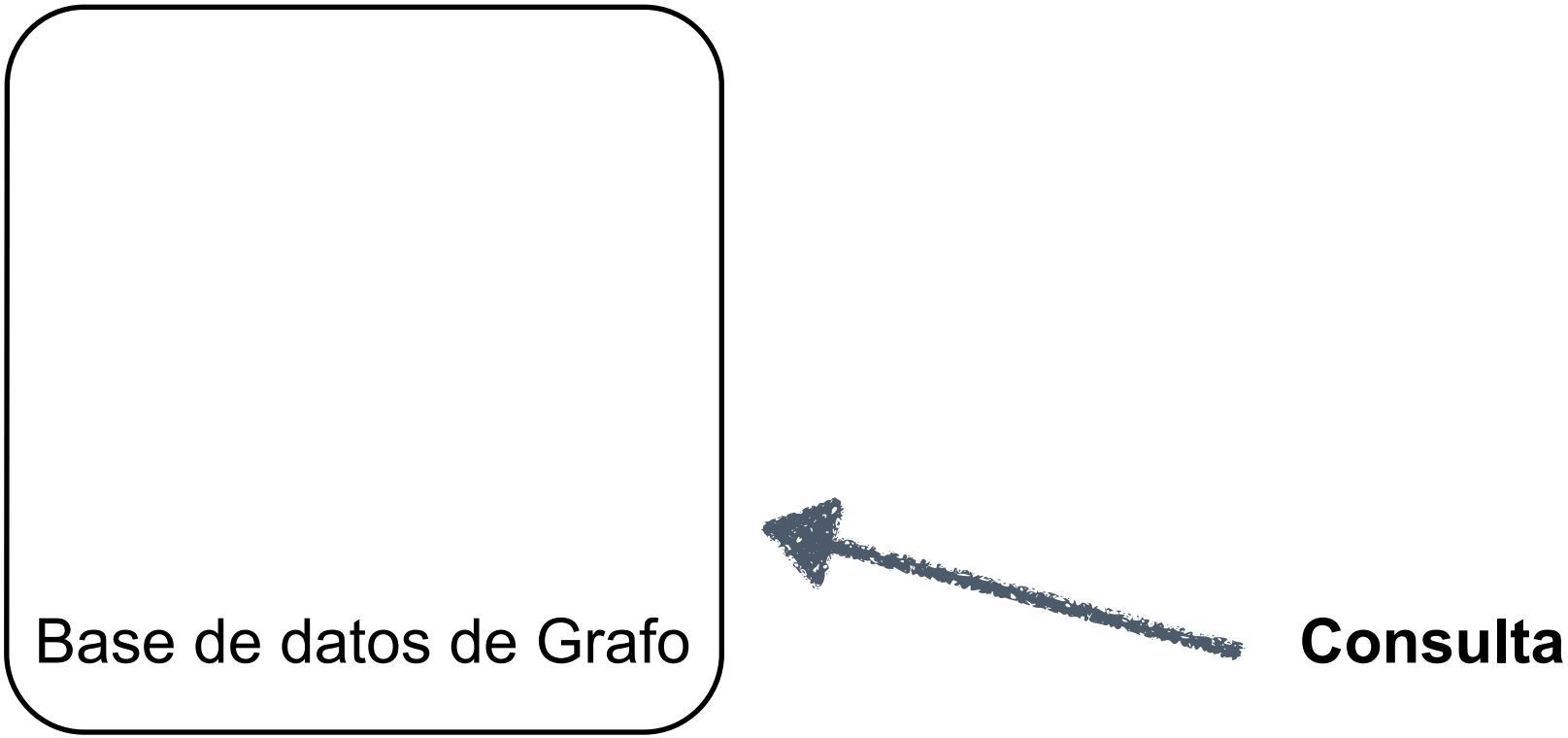
Idea:

Resumamos un grafo usando un patrón complejo

- Extraemos del grafo solo la información que es relevante
- Como si fuera una vista del grafo
- Luego dediquémonos a consultar estos patrones

En vez de acceder al grafo, resumámoslo como un patrón

En vez de acceder al grafo, resumámoslo como un patrón

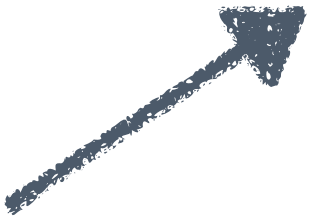


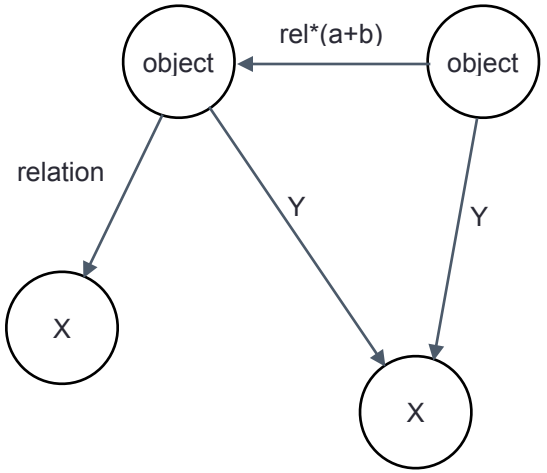
En vez de acceder al grafo, resumámoslo como un patrón

Base de datos de Grafo

Consulta

En vez de acceder al grafo, resumámoslo como un patrón





Patrón

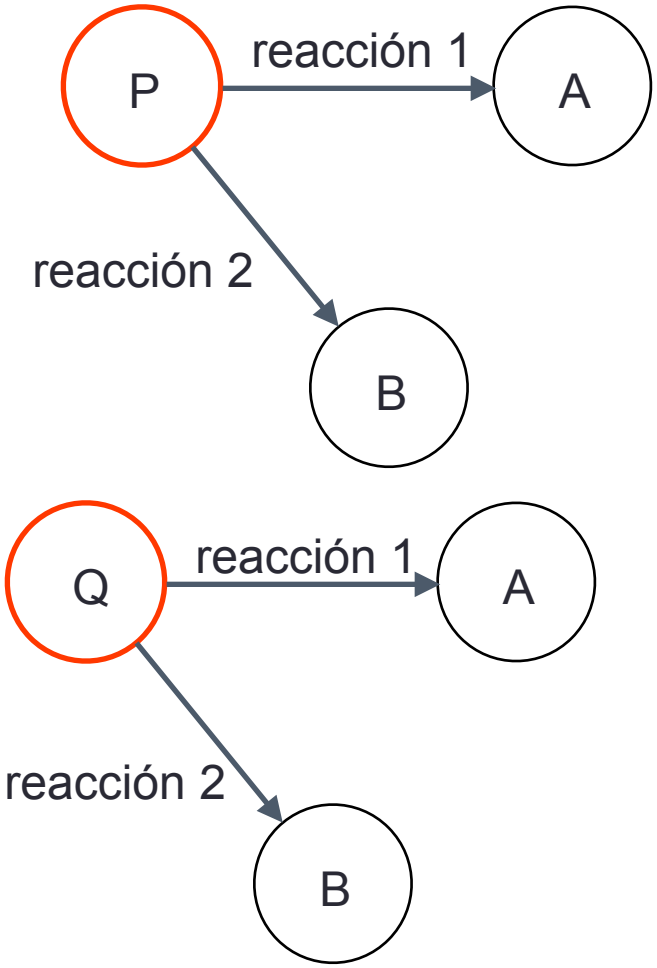
Consulta

Estudiamos patrones basados en CRPQs

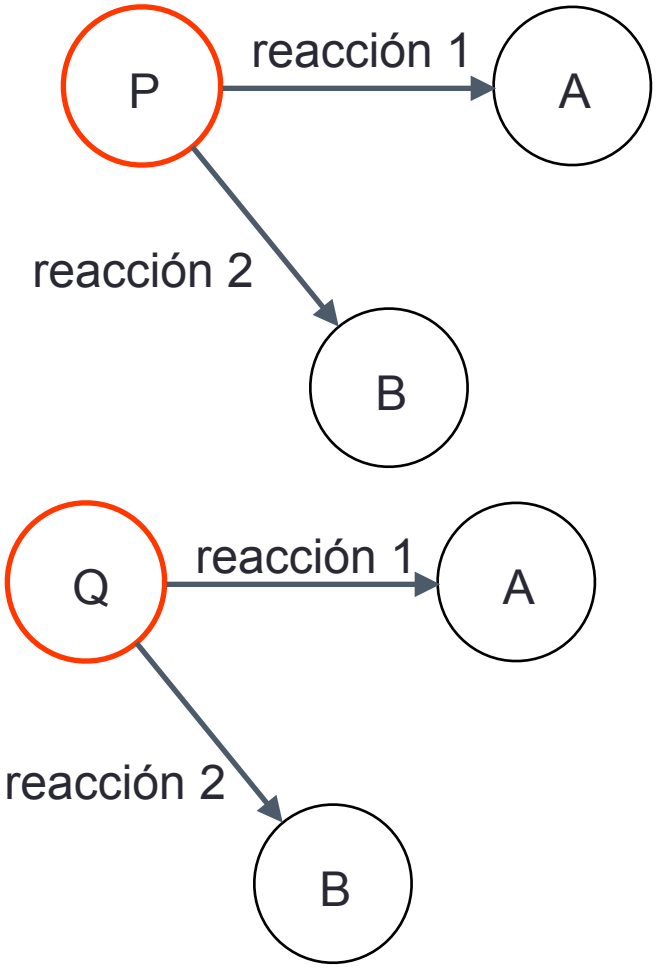
Características importantes:

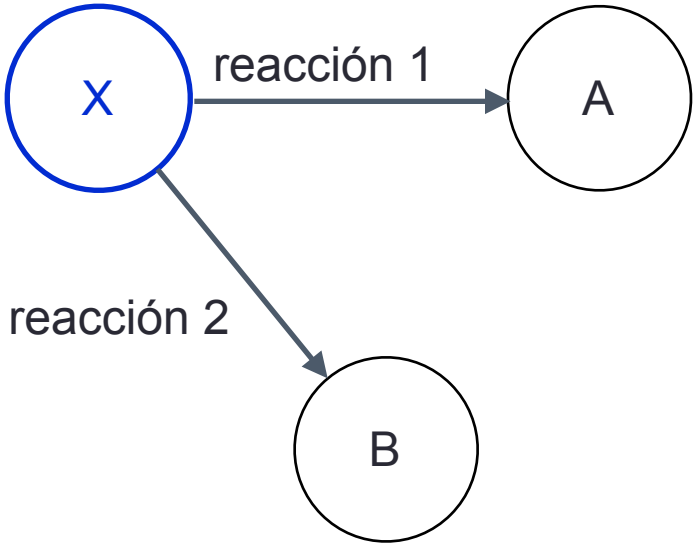
- **Variables en los nodos** para representar objetos con las mismas propiedades

Variables en los nodos
para representar objetos con las mismas propiedades



Variables en los nodos
para representar objetos con las mismas propiedades



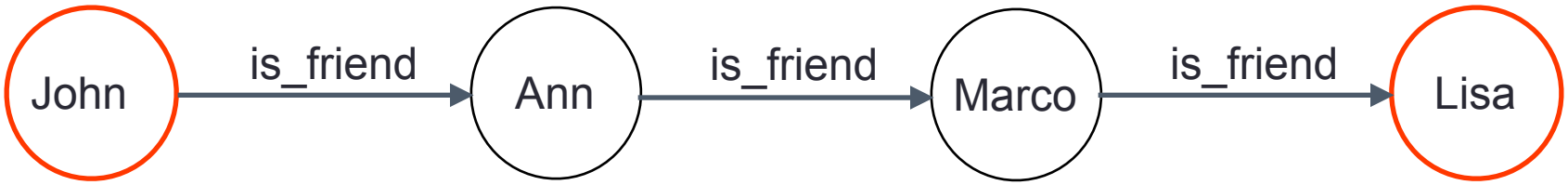


Estudiamos patrones basados en CRPQs

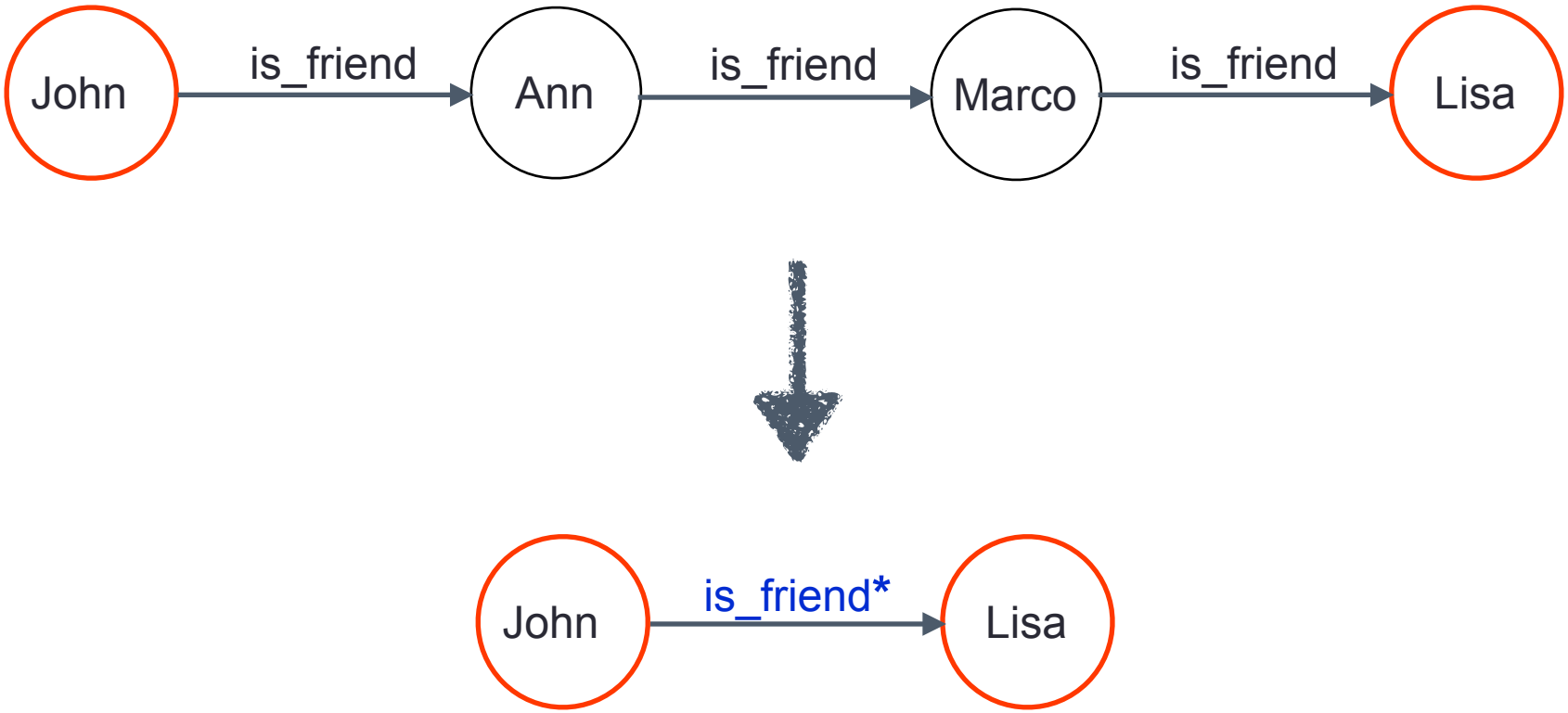
Características importantes:

- **Variables en los nodos** para representar objetos con las mismas propiedades
- **Expresiones Regulares** para representar caminos

Expresiones regulares para representar caminos



Expresiones regulares para representar caminos

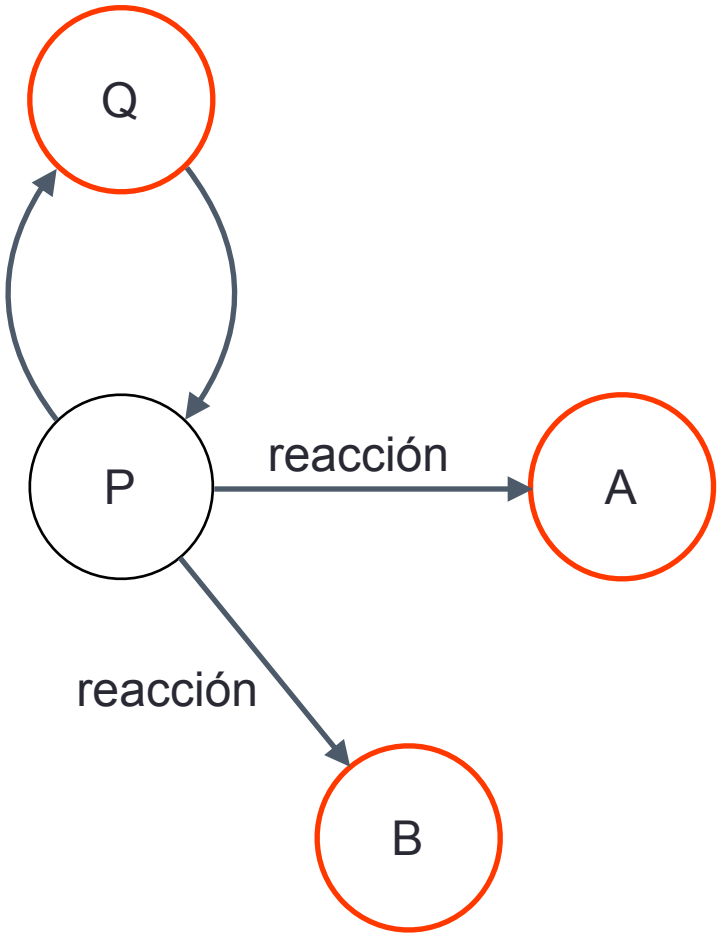


Estudiamos patrones basados en CRPQs

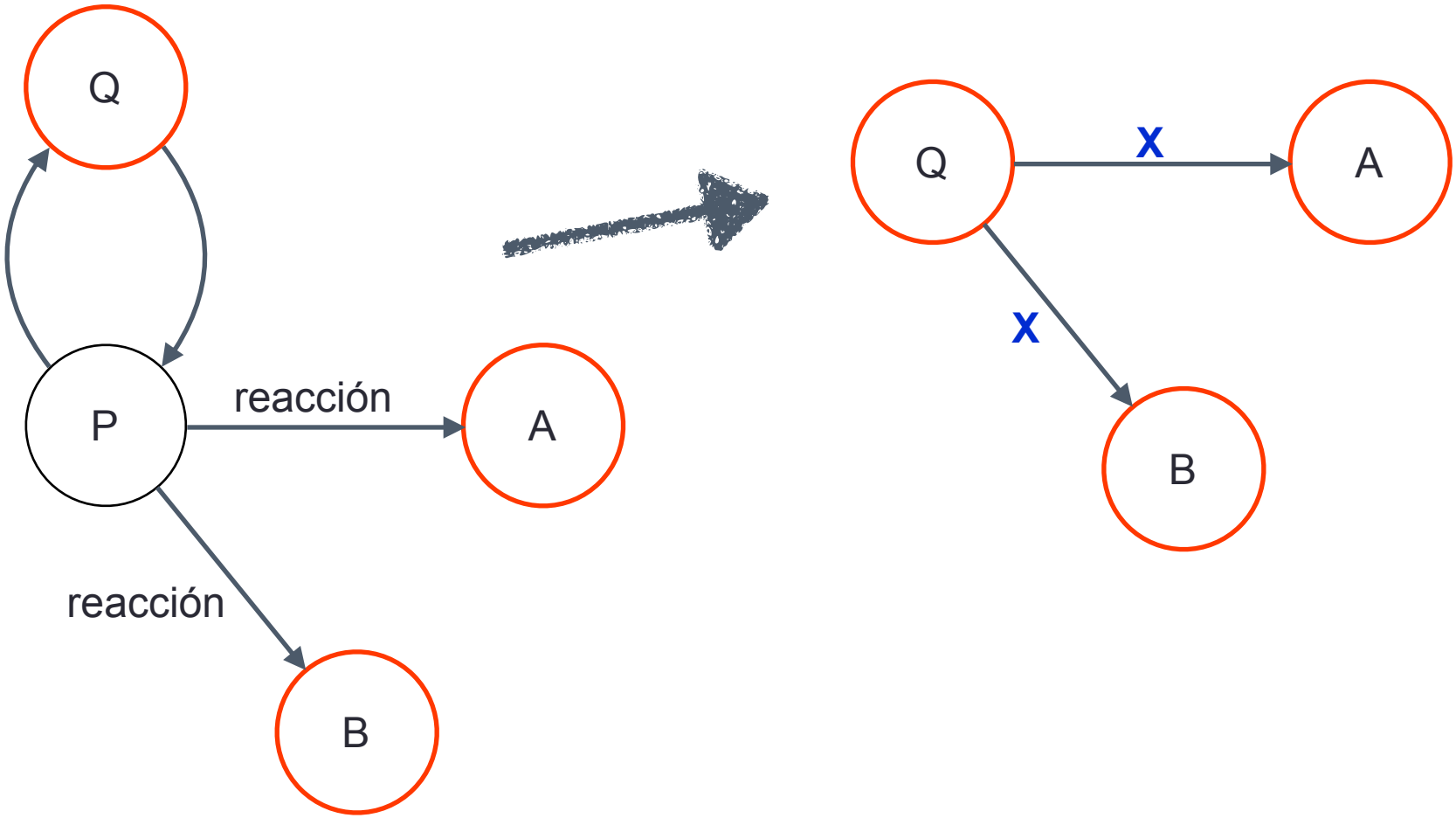
Características importantes:

- **Variables en los nodos** para representar objetos con las mismas propiedades
- **Expresiones Regulares** para representar caminos
- **Variables en las aristas** para representar relaciones con las mismas propiedades

Variables en las aristas
para representar relaciones con las mismas propiedades



Variables en las aristas
para representar relaciones con las mismas propiedades



Estudiamos patrones basados en CRPQs

Características importantes:

- **Variables en los nodos** para representar objetos con las mismas propiedades
- **Expresiones Regulares** para representar caminos
- **Variables en las aristas** para representar relaciones con las mismas propiedades

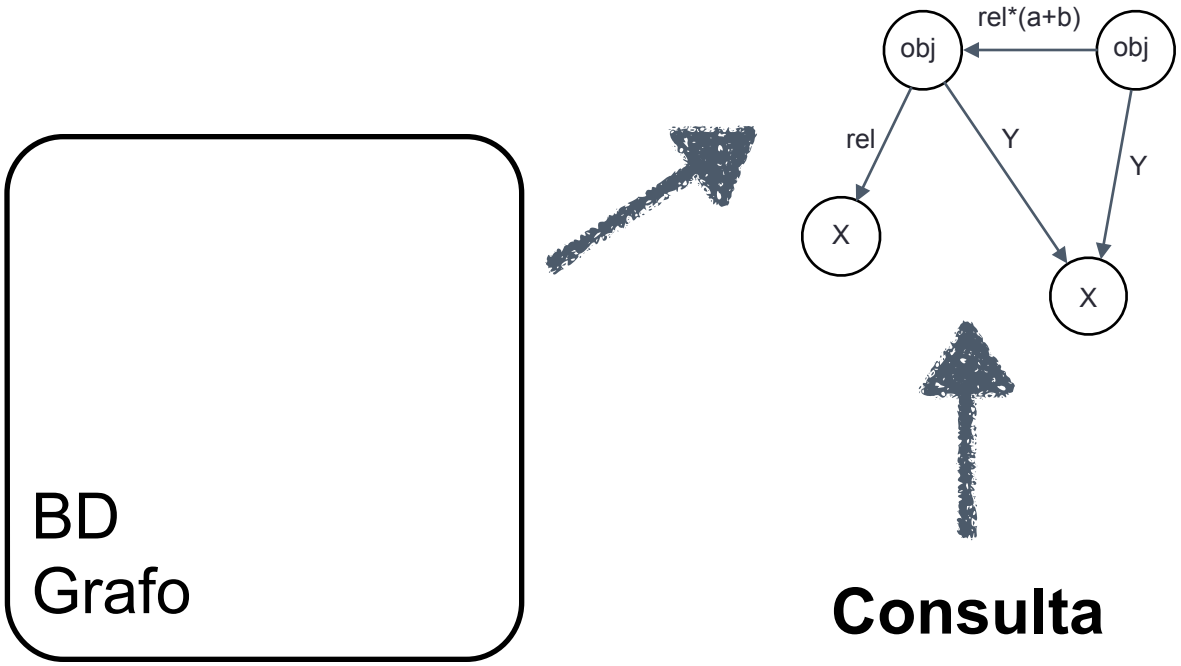
Estudiamos patrones basados en CRPQs

Características importantes:

- **Variables en los nodos** para representar objetos con las mismas propiedades
- **Expresiones Regulares** para representar caminos
- **Variables en las aristas** para representar relaciones con las mismas propiedades

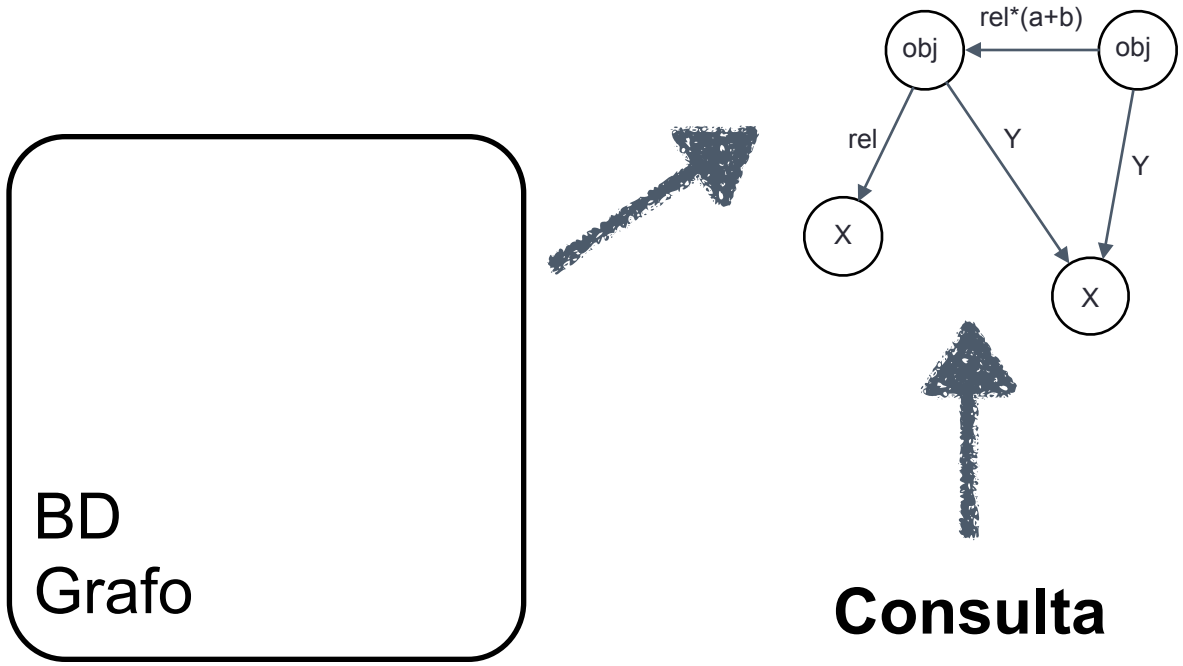
Semantica: Patrones representan varios grafos al mismo tiempo

Saber consultar patrones de grafos me permite:



Saber consultar patrones de grafos me permite:

- Extraer información de grafos resumidos
- Lidiar con información incompleta en grafos
- Transformar or integrar varios grafos



Por dónde partir?

Definir la semántica de patrones, y como consultarlos:

Por dónde partir?

Definir la semántica de patrones, y como consultarlos:

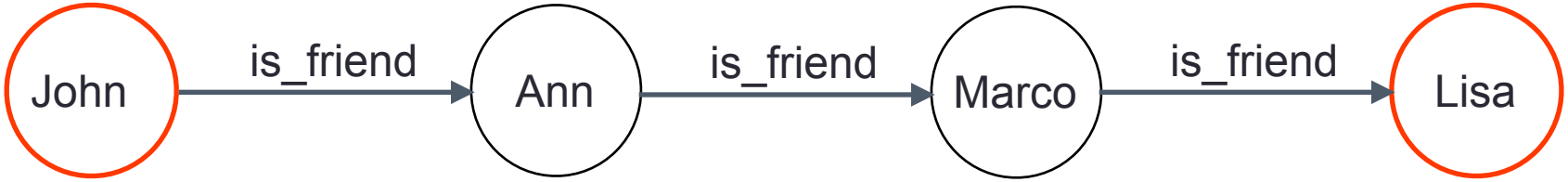
- Cada patrón representa un conjunto de BD de grafos
- Las consultas solo pueden extraer la información común a todos esos grafos

Definir la semántica de patrones, y como consultarlos:

- Cada patrón representa un conjunto de BD de grafos
- Las consultas solo pueden extraer la información común a todos esos grafos

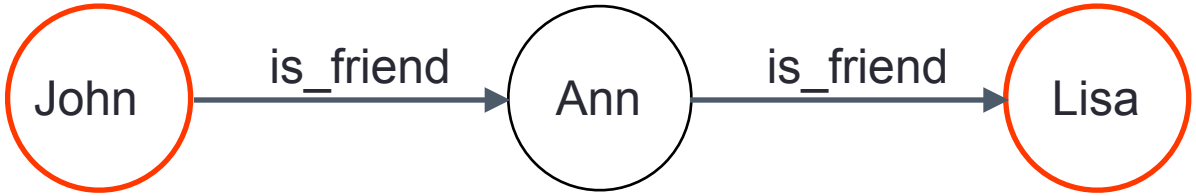
Definir la semántica de patrones, y como consultarlos:

- Cada patrón representa un conjunto de BD de grafos
- Las consultas solo pueden extraer la información común a todos esos grafos



Definir la semántica de patrones, y como consultarlos:

- Cada patrón representa un conjunto de BD de grafos
- Las consultas solo pueden extraer la información común a todos esos grafos



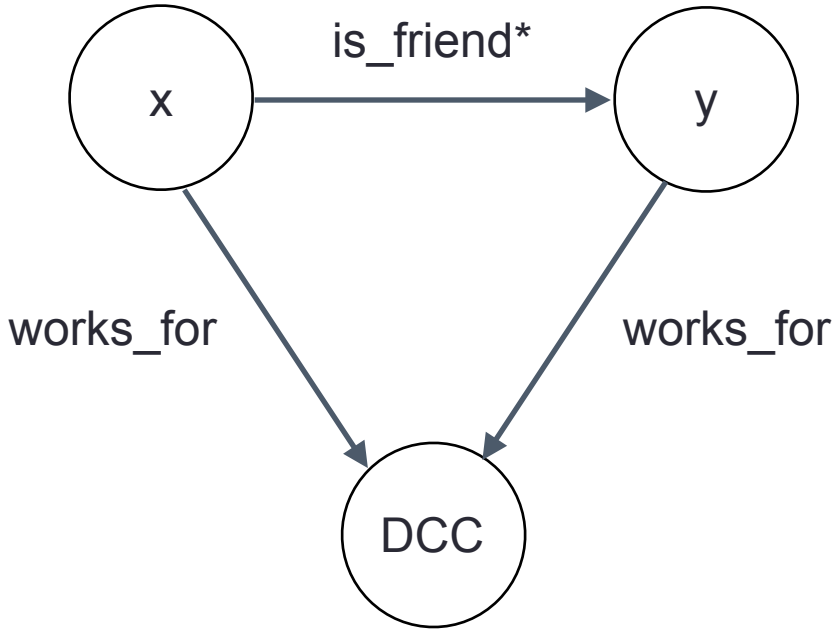
Definir la semántica de patrones, y como consultarlos:

- Cada patrón representa un conjunto de BD de grafos
- Las consultas solo pueden extraer la información común a todos esos grafos

La información común es que Juan está conectado a Lisa, via un camino de is_friend

Patrones v/s CRPQs?

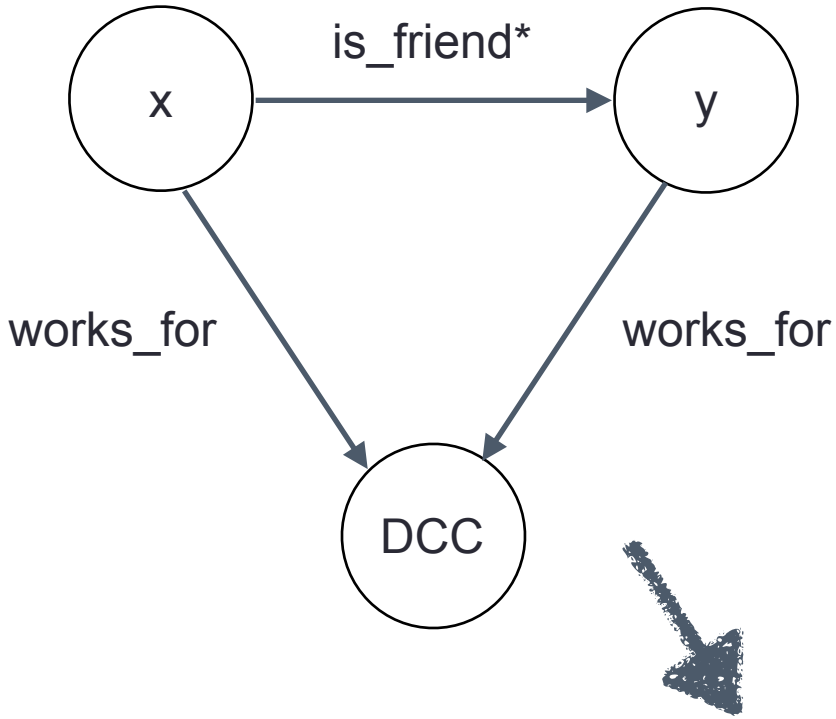
- Cada CRPQ es un patrón
- Cada patrón es una CRPQ (con variables en las aristas)



Consulta

Representación
de varios grafos

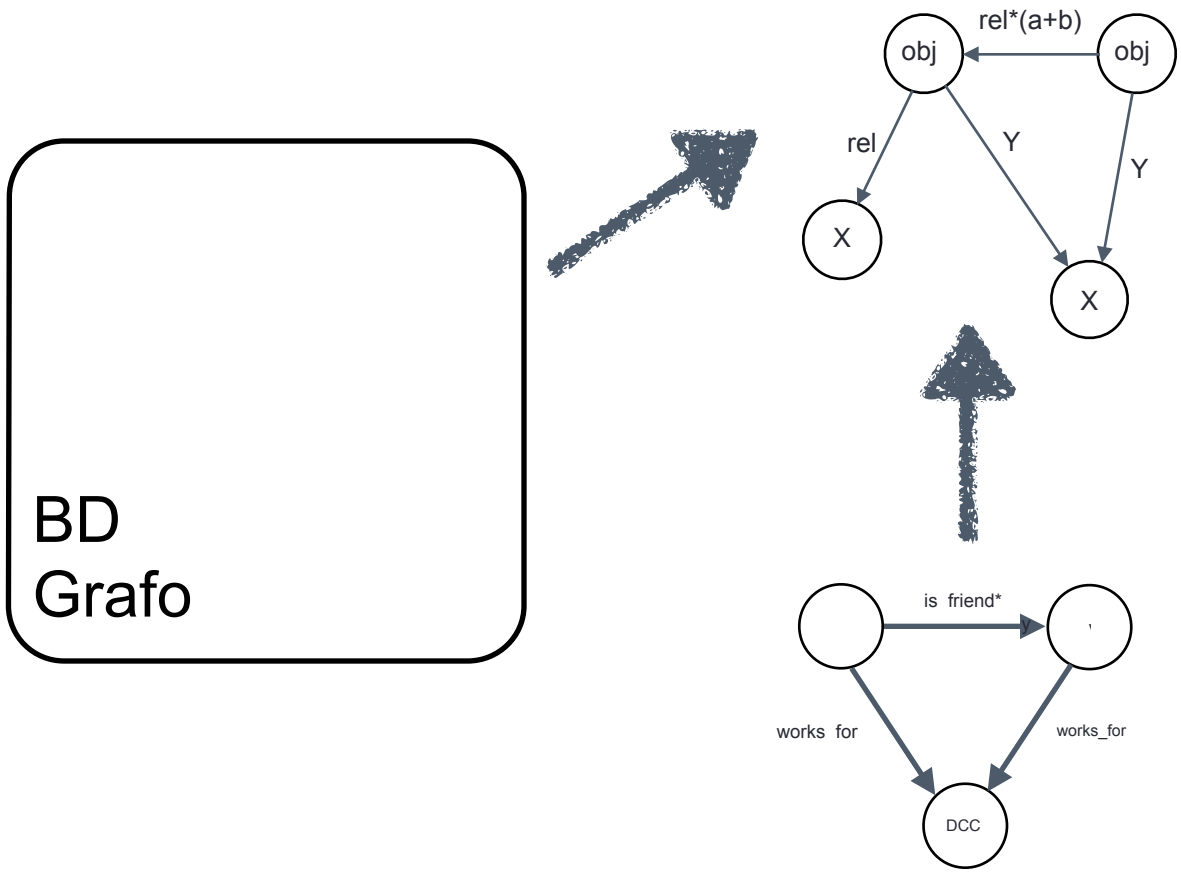
Patrones v/s CRPQs?



Consulta

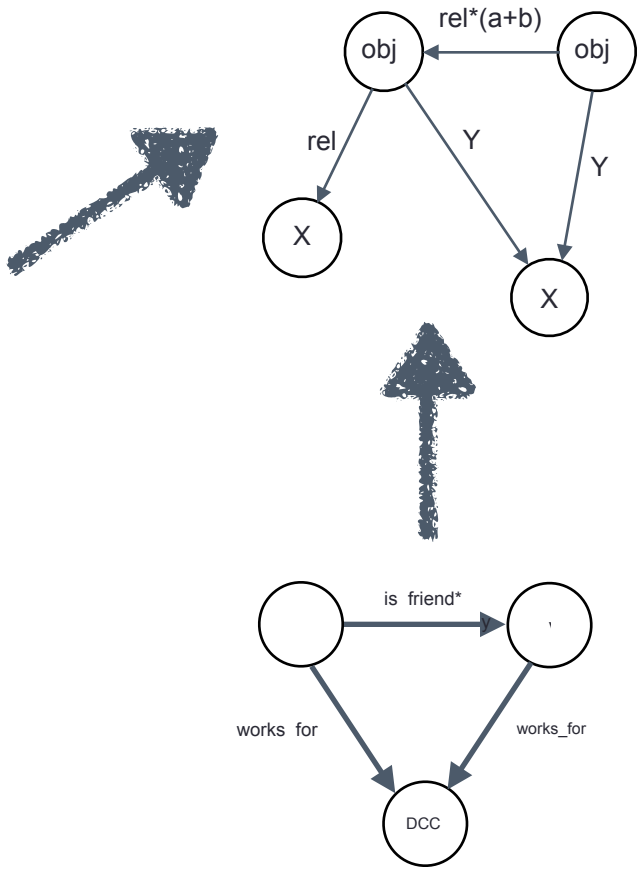
Representa todos los grafos en los que esta consulta es no vacía (modelos de esta consulta)

Incluso podemos consultar patrones…
usando patrones!



Incluso podemos consultar patrones…
usando patrones!

BD
Grafo



¿Cómo resolver este probema?

Conclusiones en cuanto a complejidad

Sean Q y G patrones.
Las respuestas de evaluar Q sobre G:

Intersección de Q sobre cada grafo representado por G

Conclusiones en cuanto a complejidad

Sean Q y G patrones.
Las respuestas de evaluar Q sobre G:

Intersección de Q sobre cada grafo representado por G

- Problema difícil (EXPSPACE)
- Si asumimos que Q está fijo, el problema es NP-completo

Conclusiones en cuanto a complejidad

Balance entre expresividad de los patrones
y complejidad de consultarlos

Conclusiones en cuanto a complejidad

Balance entre expresividad de los patrones
y complejidad de consultarlos

- Consultar de vuelve más difícil mientras los patrones son más expresivos
- Si queremos especificar caminos en los patrones,
consultar es muy costoso

Conclusiones en cuanto a complejidad

Balance entre expresividad de los patrones
y complejidad de consultarlos

Igual lo podemos hacer

- Teoría de autómatas nos da heurísticas
- Identificamos islas de eficiencia

Conclusiones en cuanto a complejidad

Balance entre expresividad de los patrones
y complejidad de consultarlos

Igual lo podemos hacer

- Teoría de autómatas nos da heurísticas
- Identificamos islas de eficiencia
- Conexiones con Constraint Satisfaction Problem para implementaciones prácticas

Maquinaria Técnica

Autómatas incompletos

- Automatas finitos con sus transiciones definidas parcialmente

Maquinaria Técnica

Autómatas incompletos

- Automatas finitos con sus transiciones definidas parcialmente
- Problemas de bases de datos de grafos se transforman en problemas estándar para estos autómatas
- Aplicaciones más allá de bases de datos
  (análisis y verificación de programas)

Resumiendo…

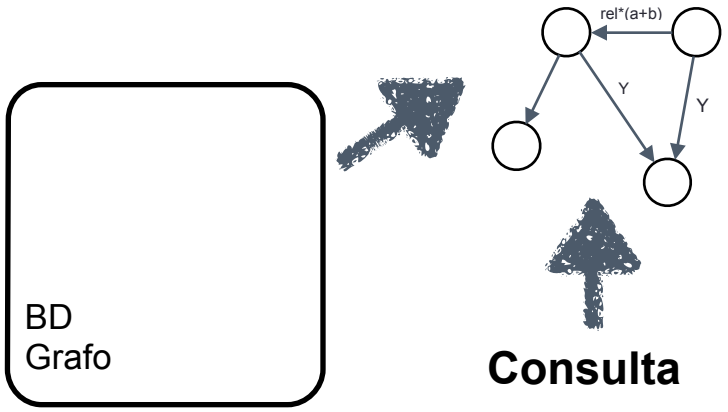
Resumiendo…

- Aplicaciones demandan lenguajes que extraigan caminos de las bases de datos de grafos
- RPQs, CRPQs,…

Resumiendo…

Sabemos como consultar

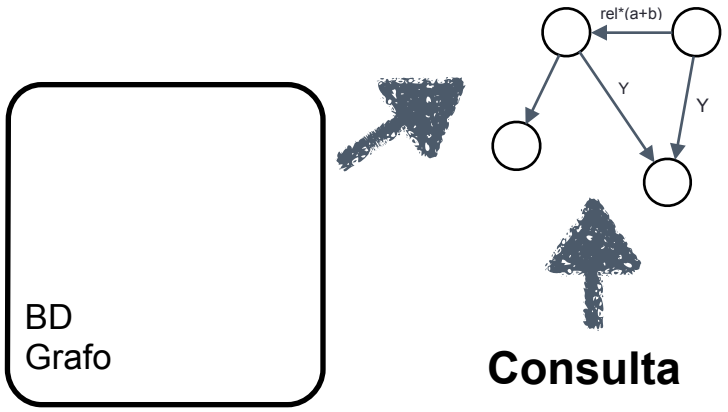
patrones de grafos complejos



Resumiendo…

Sabemos como consultar

patrones de grafos complejos



- Identificamos la complejidad de este problema
- Semántica de patrones para representar bases de datos
- Modelo Teórico con aplicaciones en otras áreas

Hacia adonde apuntamos

Hacia adonde apuntamos

Sabemos como consultar patrones, pero como construirlos?

- Dado un grafo, crear el patrón (pequeño) mas representativo

Hacia adonde apuntamos

Sabemos como consultar patrones, pero como construirlos?

- Dado un grafo, crear el patrón (pequeño) mas representativo

Ayudando a los humanos a entender las BD de grafos

- DBpedia: las páginas se almacenan de la misma forma
- tal vez podemos representar esto como un patrón
  (quizá con características adicionales)