

Minería Semántica y Minería de Procesos para automatización

Jose Aguilar

Minería de Procesos

Jose Aguilar

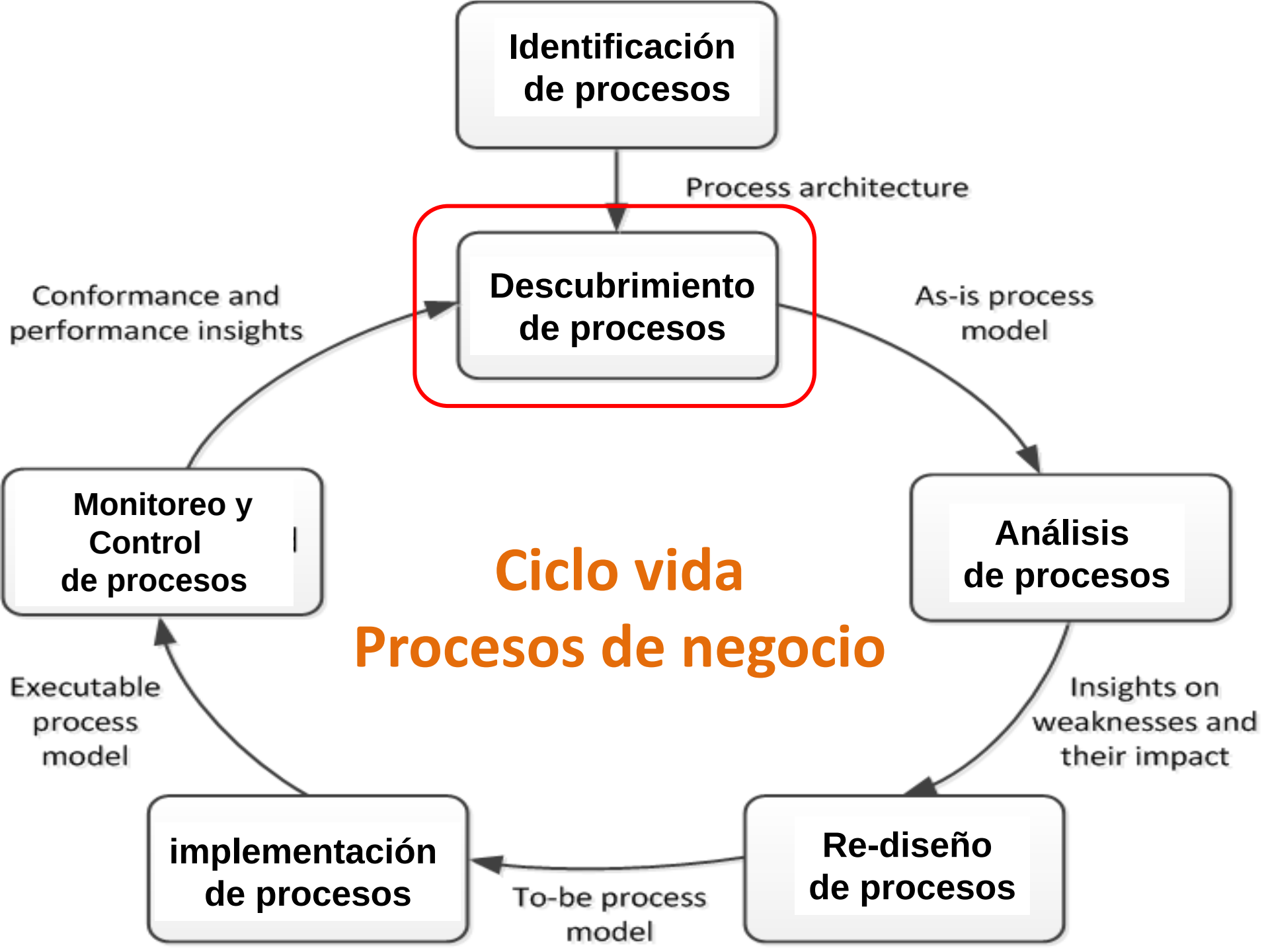
Everything Mining

Minería de Datos	Minería de Procesos
Orientado por los Datos	Orientado por los Procesos
Utiliza datos reales de los procesos	Usa el registro de eventos de procesos
Respuestas a preguntas relacionadas con datos	Respuestas a preguntas relacionadas con datos procesos.
El modelo está entrenado para hacer predicciones, buscar patrones	Descubre o crea el modelo de proceso completo y real.
Normalmente se usa para predecir situaciones.	Se usa para detectar fallas, mejorar el proceso, etc.

Minería de Procesos

- Comprender y mejorar los "flujos de atención" en un hospital mediante el uso inteligente de eventos de datos dispersos en cientos de tablas de bases de datos con los datos del paciente.





Minería de Procesos

Casos de uso de minería de procesos

- ¿Cuál es el proceso que la gente realmente sigue?
- ¿Dónde están los cuellos de botella en mi proceso?
- ¿Dónde se desvían las personas (o las máquinas) del proceso esperado o idealizado?

Minería de Procesos

Casos de uso de minería de procesos

- ¿Cuál es el proceso que la gente realmente sigue?
- ¿Dónde están los cuellos de botella en mi proceso?
- ¿Dónde se desvían las personas (o las máquinas) del proceso esperado o idealizado?

Minería de Procesos

Casos de uso de minería de procesos

- ¿Cuáles son las "carreteras" en mi proceso?
- ¿Qué factores están influyendo en un cuello de botella?
- ¿Podemos predecir problemas (retraso, desviación, riesgo, etc.) para casos en ejecución?
- ¿Podemos recomendar contramedidas?
- ¿Cómo rediseñar el proceso / organización / máquina?
- ...

Minería de Procesos

Premisas de M. Hammer

- Cualquier proceso es mejor que nada
- Un buen proceso es mejor que un mal proceso
- Cualquier buen proceso se puede mejorar
- Todo buen proceso en el tiempo se vuelve mal proceso, sin intención

Minería de Procesos

Ámbitos de la minería de procesos

- Preguntas, problemas y soluciones orientadas al rendimiento,
- Análisis del modelo de proceso (simulación, verificación, optimización, juegos, etc.)
- Análisis orientado al datos (minería de datos, aprendizaje automático, inteligencia de negocio)
- Preguntas, problemas y soluciones orientados al cumplimiento

Minería de Procesos

EVENTOS

Id evento

Nombre evento

marca de tiempo

student name	course name	exam date	mark
Peter Jones	Business Information systems	16-1-2014	8
SandyScott	BusinessInformation systems	16-1-2014	
Bridget White	BusinessInformation systems	16-1-2014	
John Anderson	BusinessInformation systems	16-1-2014	
SandyScott	BPM Systems	17-1-2014	7
Bridget White	BPM Systems	17-1-2014	8
SandyScott	Process Mining	20-1-2014	5
Bridget White	Process Mining	20-1-2014	9
John Anderson	Process Mining	20-1-2014	8
...	...	...	...

cada fila es un evento (aquí: un examen)

Minería de Procesos

EVENTOS

Id evento

Nombre evento

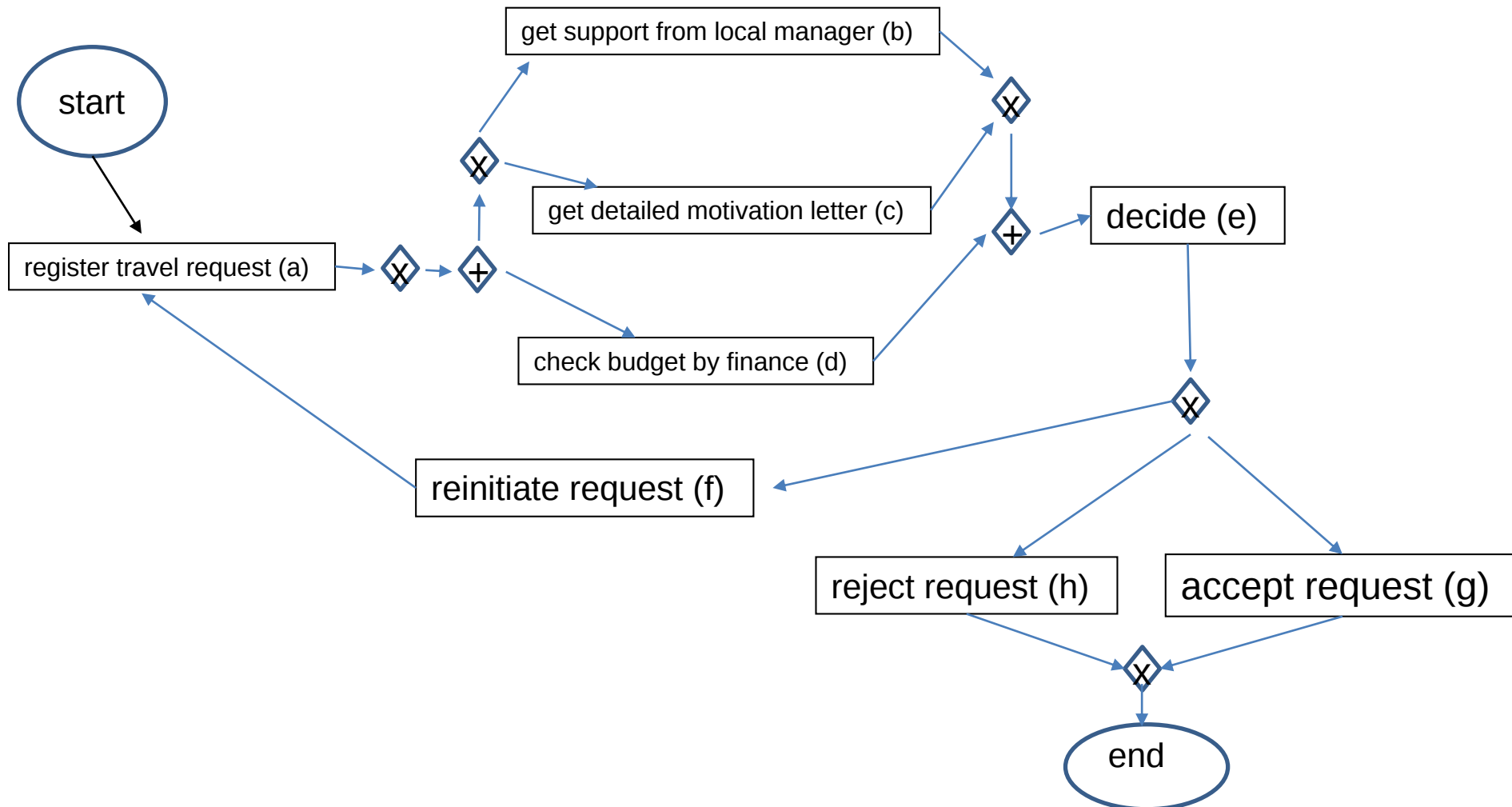
marca de tiempo

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	bloodtest	23-1-2014@10.18	Dr. Scott	61	40.00
5833	bloodtest	23-1-2014@10.27	Dr. Scott	24	40.00
5781	bloodtest	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr.Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	CarolHope		450.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00
• ...	...	...	...	...	...

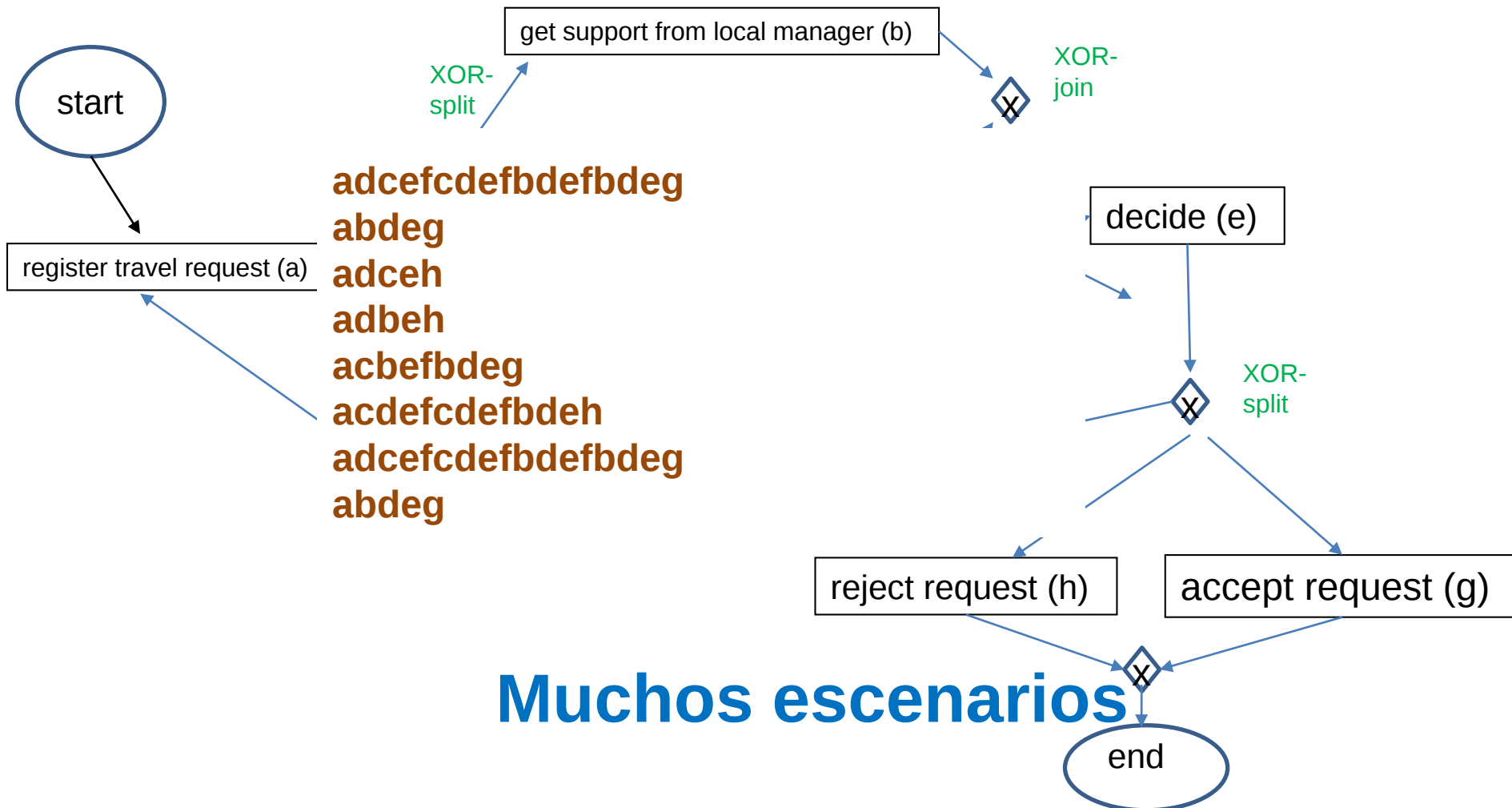
Minería de Procesos

Case	Activity	Timestamp	Resource
432	register travel request (a)	18-3-2014:9.15	John
432	get support from local manager (b)	18-3-2014:9.25	Mary
432	check budget by finance (d)	19-3-2014:8.55	John
432	decide (e)	19-3-2014:9.36	Sue
432	accept request (g)	19-3-2014:9.48	Mary

Minería de Procesos



Minería de Procesos



Minería de Procesos

adcefcdefbdefbdeg
abdeg
adceh
adbeh
acbefbdeg
acdefcdefbdeh
adcefcdefbdefbdeg
abdeg

**Crear un modelo de proceso que siga
los rastros de arriba**

Minería de Procesos

Por qué

- Entender los procesos como realmente son y no como uno se lo podría imaginar
- Respalda hipótesis con evidencias, y no con intuiciones o creencias
- Cuantificar impacto de un rediseño: antes y después

Minería de Procesos

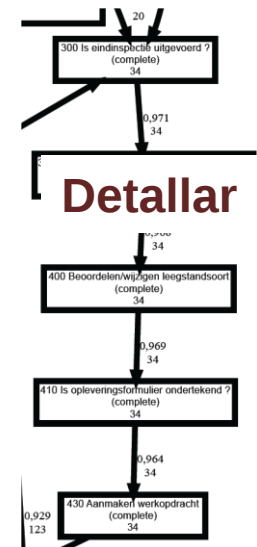
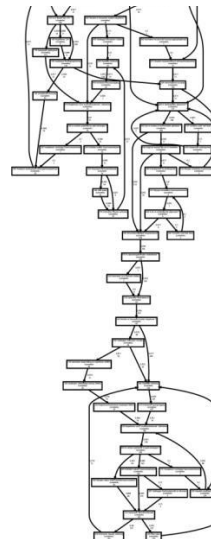
Ejemplo de descubrimiento de proceso (agencia de viaje, 1810 casos, 15877 eventos)

117315	110 Bepalen leegstandsoort	16.05.2007 14:06:23
117315	120 Plannen eindinspectie	16.05.2007 14:36:01
117315	130 Is het opleveringsformulier ondertekend ?	23.05.2007 09:41:40
117315	150 Is er sprake van ZAV ?	23.05.2007 09:41:51
117315	170 Aanpassen plattegrond	23.05.2007 11:57:18
117315	180 Aanpassen woningwaardering	23.05.2007 09:42:37
117315	190 Actualiseren huurprijs	23.05.2007 09:48:23
117315	200 Toewijzen woning/bedr.ruimte/gar/berg/park/ops	23.05.2007 09:48:29
117315	210 Registreren voorl. huurovereenkomst +afdrukken	10.09.2007 16:24:36
117315	22	56:18
117315	24	17:12
117315	25	39:59
117315	26	51:24
117315	27	52:08
117315	30	47:04
117315	340 Zijn er nieuwe of niet herstelde gebreken ?	07.06.2007 14:47:06
117315	400 Beoordelen/wijzigen leegstandsoort	07.06.2007 14:51:16
117315	410 Is opleveringsformulier ondertekend ?	07.06.2007 14:51:26
117315	430 Aanmaken werkopdracht	11.06.2007 09:21:39
117315	440 Worden er bonussen/ kosten toegekend ?	11.06.2007 09:21:49
117315	460 Opstellen eindnota	08.08.2007 16:18:26
117315	470 Archiveren huuropzegging	09.08.2007 14:42:23
119763	010 Registreren huuropzegging	09.05.2007 11:19:14
119763	030 Vastleggen toekomstige adres	09.05.2007 12:25:01
119763	050 Inplannen afspraak 1e inspectie	09.05.2007 11:59:52
119763	060 Aanmaken bevestigingsbrief / huuropzeggingform.	09.05.2007 12:31:57
119763	070 Is 1e inspectie uitgevoerd ?	16.05.2007 13:04:26
119763	100 Gereedmelden 1e insp. / Voorcalculatie maken	16.05.2007 13:43:39
119763	110 Bepalen leegstandsoort	16.05.2007 13:43:28
119763	120 Plannen eindinspectie	16.05.2007 13:42:58
119763	130 Is het opleveringsformulier ondertekend ?	16.05.2007 13:34:49
119763	150 Is er sprake van ZAV ?	16.05.2007 13:34:56

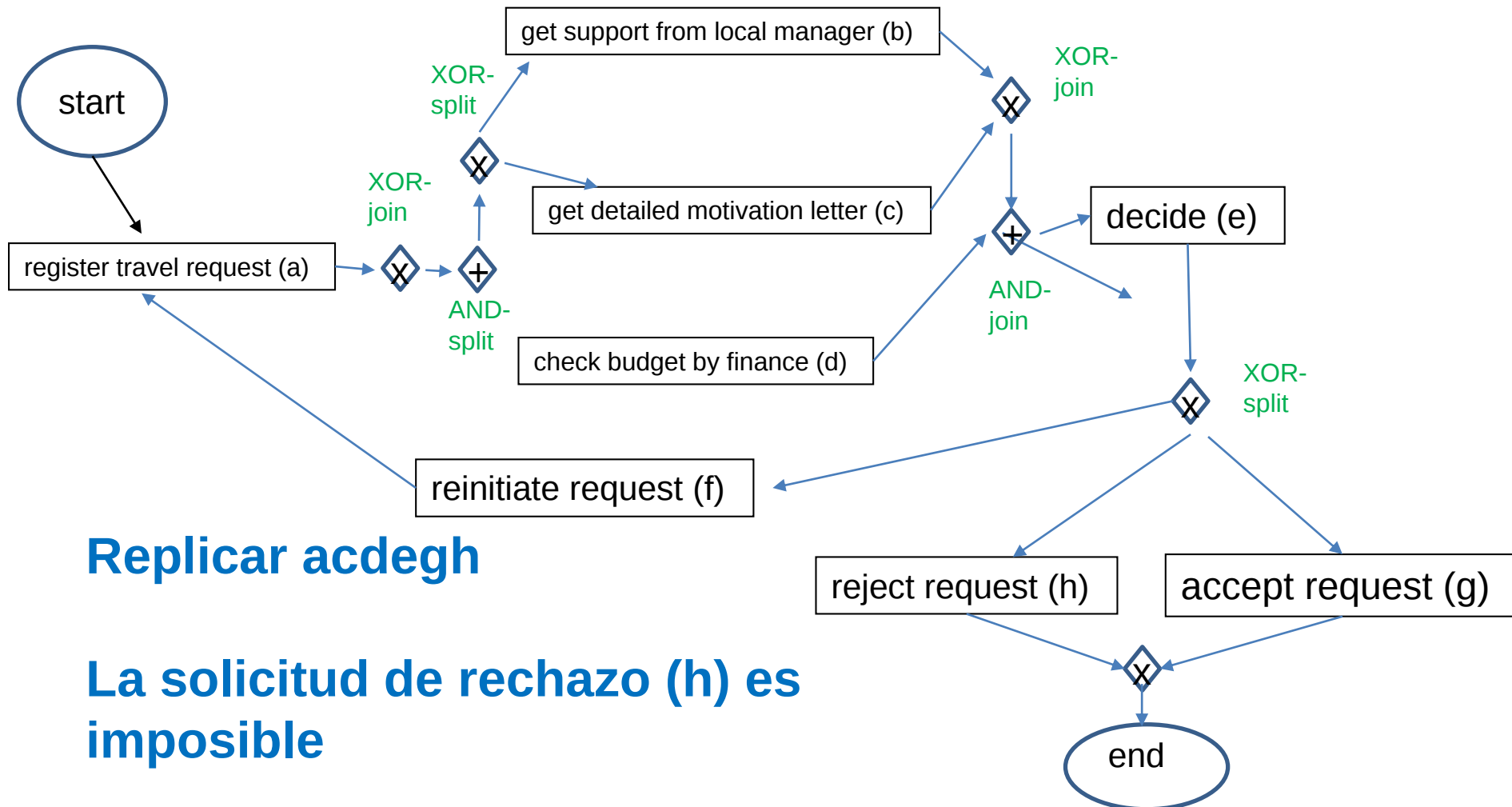
Registro de eventos En la agencia



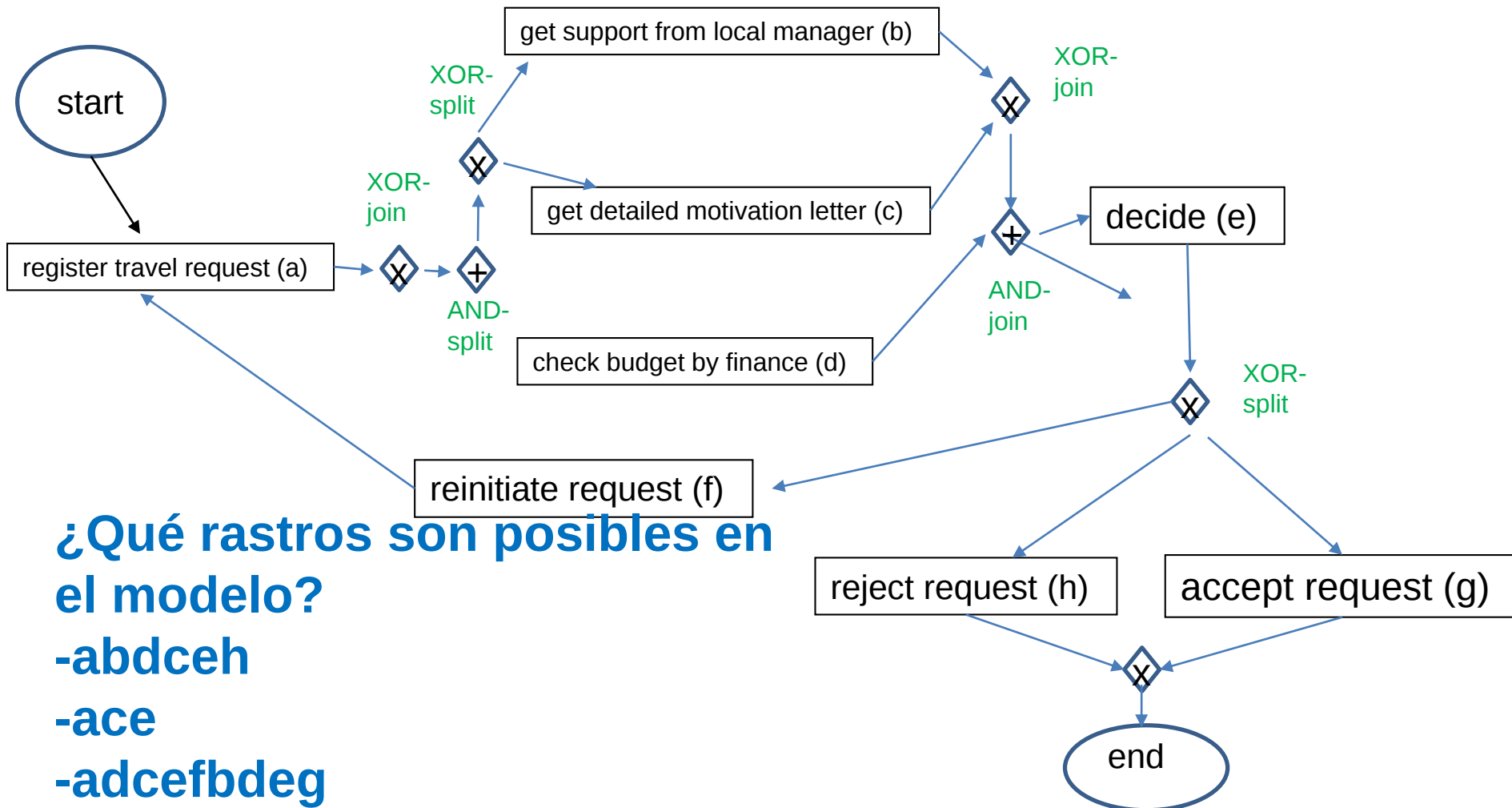
Diagrama de actividades



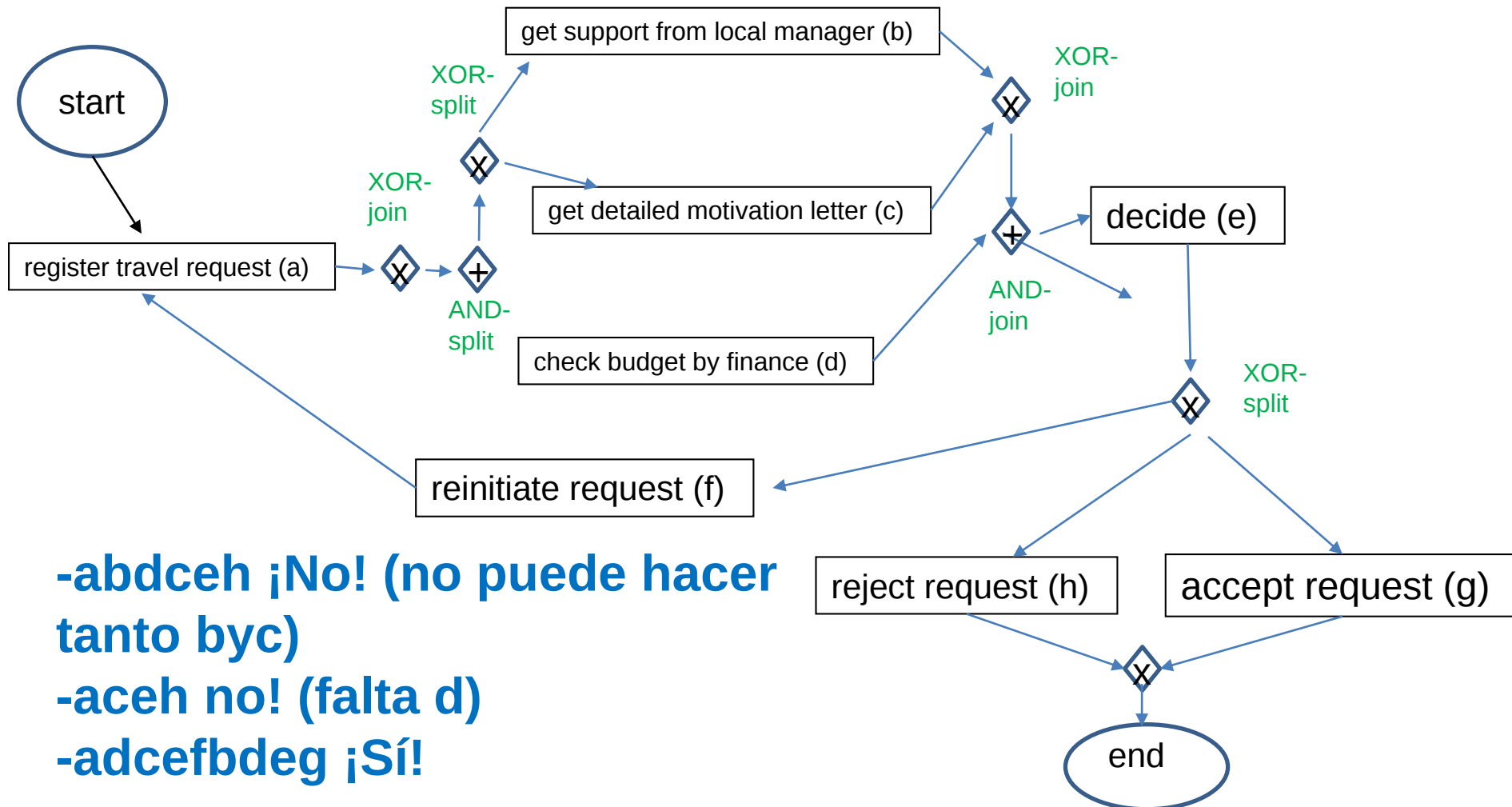
Minería de Procesos



Minería de Procesos

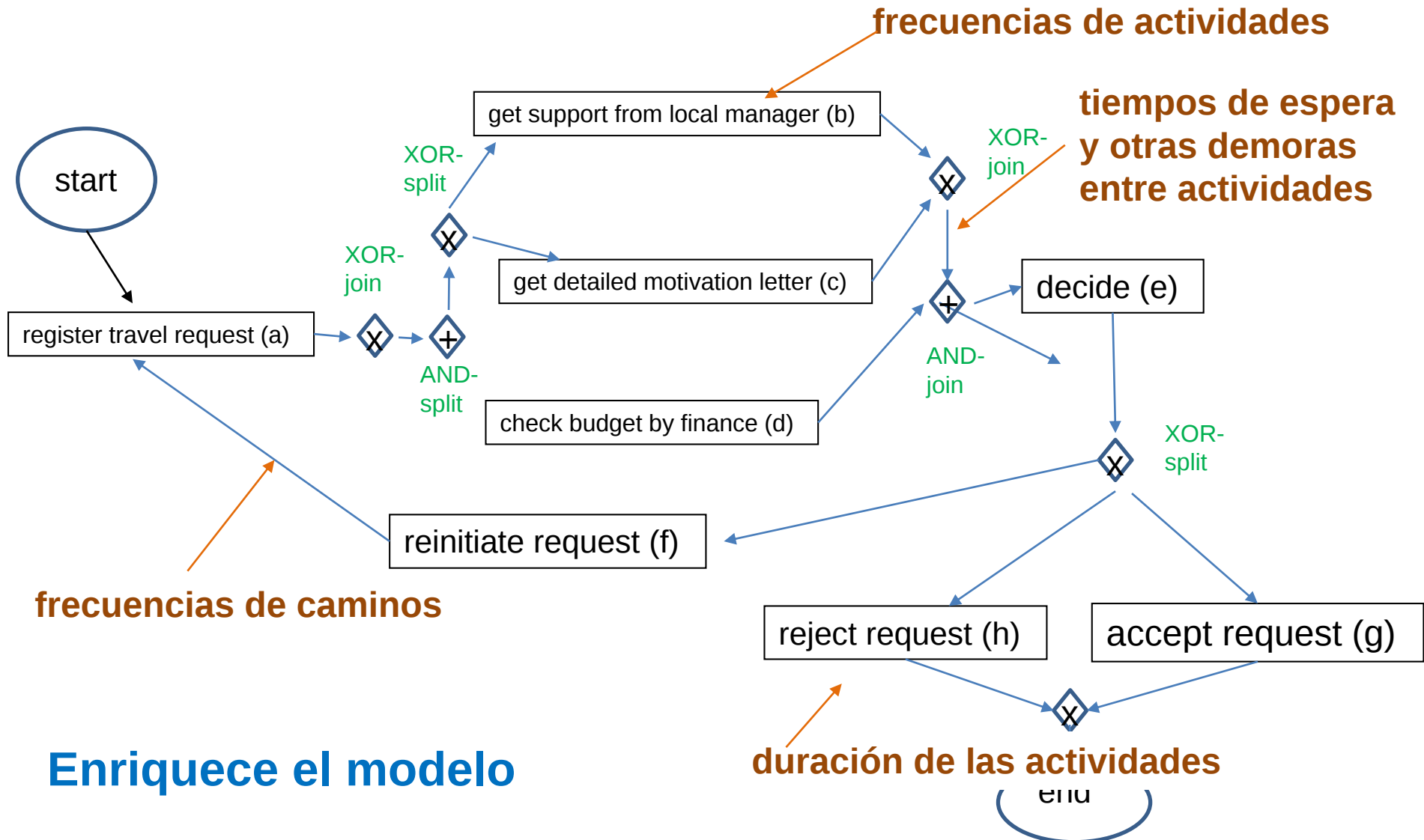


Minería de Procesos

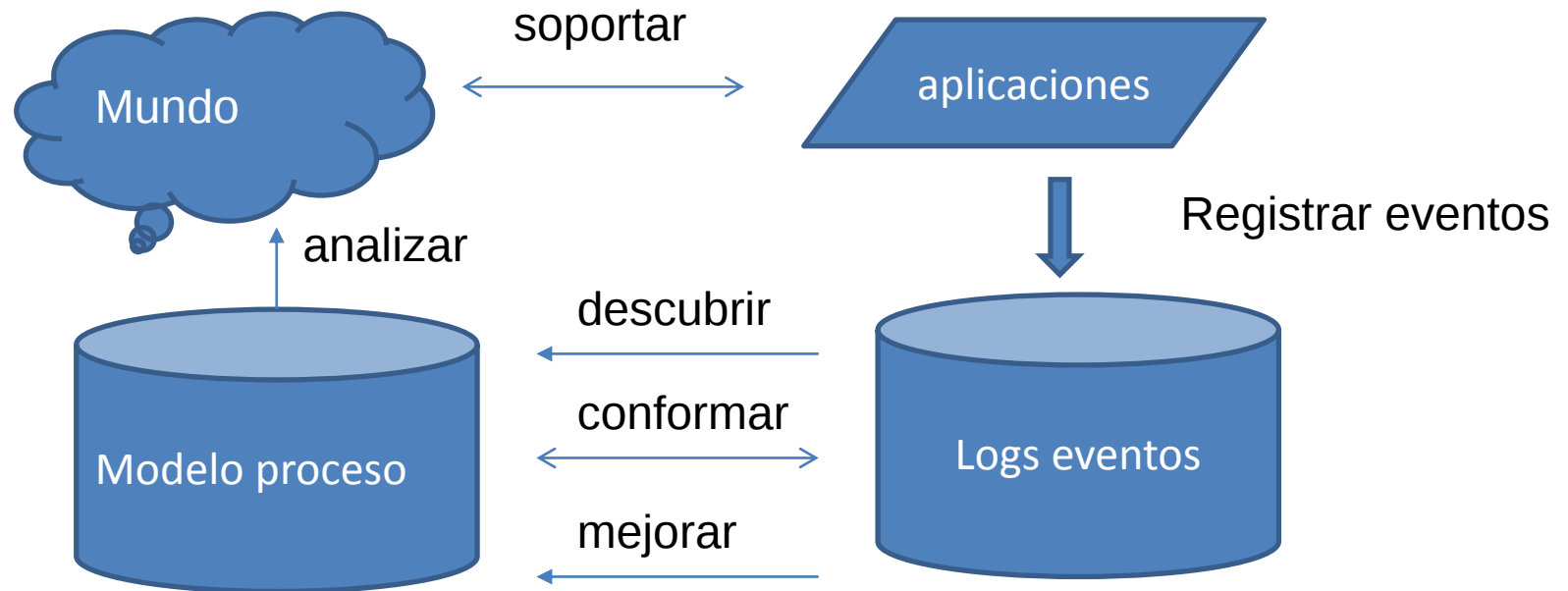


-abdceh ¡No! (no puede hacer tanto byc)
-aceh no! (falta d)
-adcefbdeg ¡Sí!

Minería de Procesos



Minería de Procesos



Crear un modelo de proceso que siga los rastros de arriba

Minería de Procesos

descubrir procesos

minar procesos

BPM

predecir procesos

conformar procesos

Minería Datos

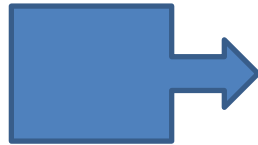
Minería de Procesos

El descubrimiento de procesos es como aprender un idioma:

abc ?

ab(c|d) ?

ab*(c|d) ?



abc

abd

ac

sentencia = trace en el registro de eventos

lenguaje = modelo del proceso

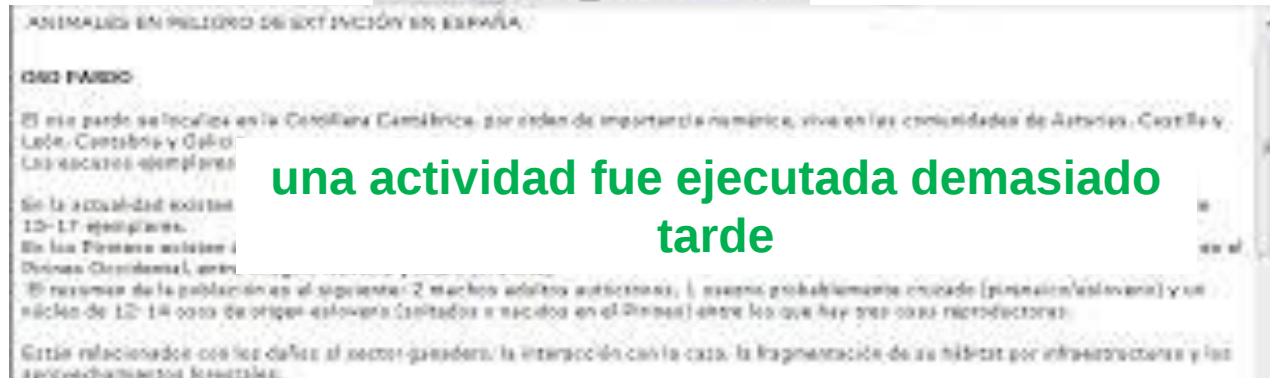
Minería de Procesos

La verificación de conformidad es como un corrector ortográfico

una actividad que no debería suceder sucedió

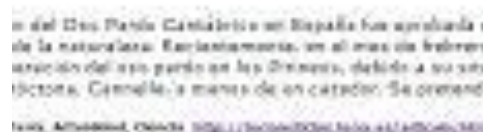


una actividad fue ejecutada por la persona equivocada



una actividad fue ejecutada demasiado tarde

una actividad que debería suceder no sucedió



algunas actividades fueron intercambiadas

Minería de Procesos

Técnicas de Interés de MD

árboles de decisión

agrupamiento

minería de e

reglas de asociación

minería de secuencia

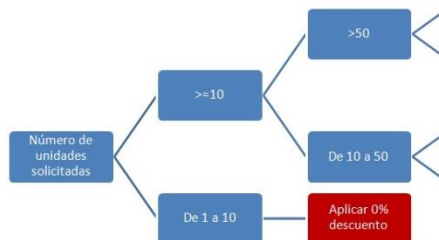
¡Las técnicas de

minería de datos no

consideran los modelos

de proceso de extremo a

extremo!



árboles de decisi...

minería de secuencia

$\langle \{vino\ rosso\}, \{pizza\ romana, pizza\ siciliana\}, \{gelato\} \rangle \Rightarrow \langle \{espresso\} \rangle$

$\{pizza\ siciliana\}$

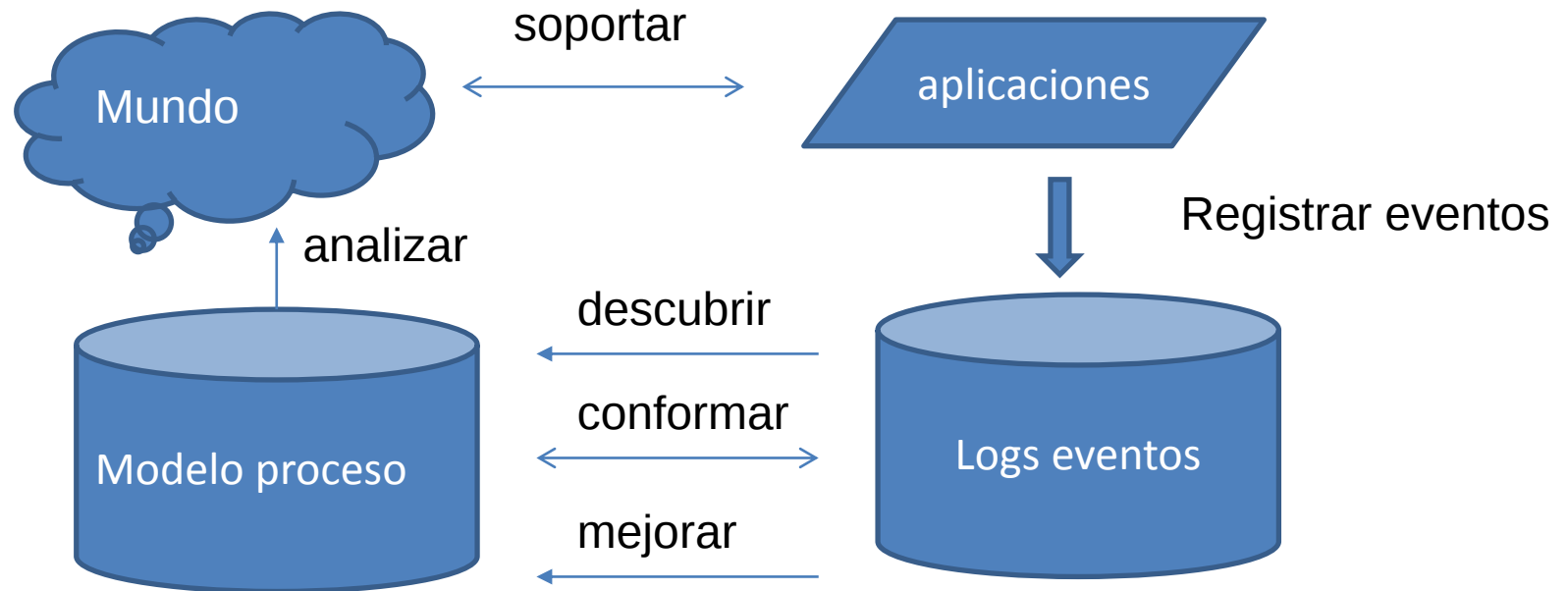
$port(R_3) = 0.354$

$snce(R_3) = 0.904$

$lift(R_3) = 1.886$

acción

Minería de Procesos



Crear un modelo de proceso que siga los rastros de arriba

Minería de Procesos

Métodos exploratorios

- Descubrir modelos
- Evaluar rendimiento sobre modelos
- Descubrir variantes

Métodos dirigidos por preguntas

- Definir un problema
- Descomponerlo en preguntas => analizar

Minería de Procesos

Métodos L*: dirigido por preguntas

- Planeación
- Extracción y pre-procesamiento
- Análisis de datos => minería
- Interpretación

Minería de Procesos

Métodos L* dirigido por preguntas

Planeación

- Determinar la pregunta
 - Identificar variables y sus variaciones (temporales, por geografía, etc.)
 - Diagnosticar rendimientos (cuellos de botella, ...)
 - Identificar causas
 - Definir métricas y criterios precisos

Minería de Procesos

Métodos L*: dirigido por preguntas

Extracción y pre-procesamiento

- Encontrar los datos
 - Sistema de información internos o Internet
 - Organizar por trazas
 - Convertir en formato estándar (XES)
- Combinación de múltiples fuentes
- Depurar los datos
 - Filtrar eventos no frecuentes
 - Filtrar variantes infrecuentes
 - Combinar eventos equivalentes

Minería de Procesos

Métodos L*: dirigido por preguntas

Análisis de datos

- Descubrir modelos
- Descubrir caminos frecuentes/infrecuentes
- Calcular métricas
 - Tiempos de ciclo, tiempos de espera, % de errores
- Descubrir clases de casos
 - Clasificación

Minería de Procesos

¿Qué puedo hacer con MD?

- Filtrado de eventos
- Agrupamiento de trazas
- Abstracción del modelo resultante
 - Tareas o caminos de interés
 - Agregación de subprocessos
- Extracción de reglas de interés

Minería de Procesos

Reglas de interés

- **Reglas de decisión**

¿Cómo se toman decisiones en diferentes puntos del proceso?

- **Reglas descriptivas**

¿ Cuándo y por qué se ejecuta una tarea?

- **Reglas discriminativas**

¿Cuándo y por qué terminamos con resultados negativos

Business Process Model and Notation (BPMN)

- Estándar OMG (hoy en día BPMN 3.0)
- Compatible con numerosas herramientas: bpmn.org enumera más de 70 herramientas
- Tanto para modelos conceptuales como ejecutables

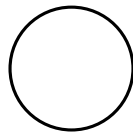


Business Process Model and Notation (BPMN)

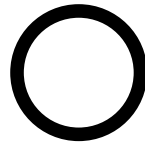
Un modelo de proceso BPMN es un gráfico que consta de cuatro tipos de elementos centrales:



actividad

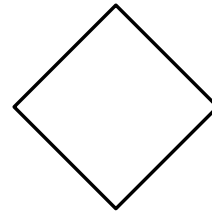


fin

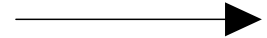


inicio

event



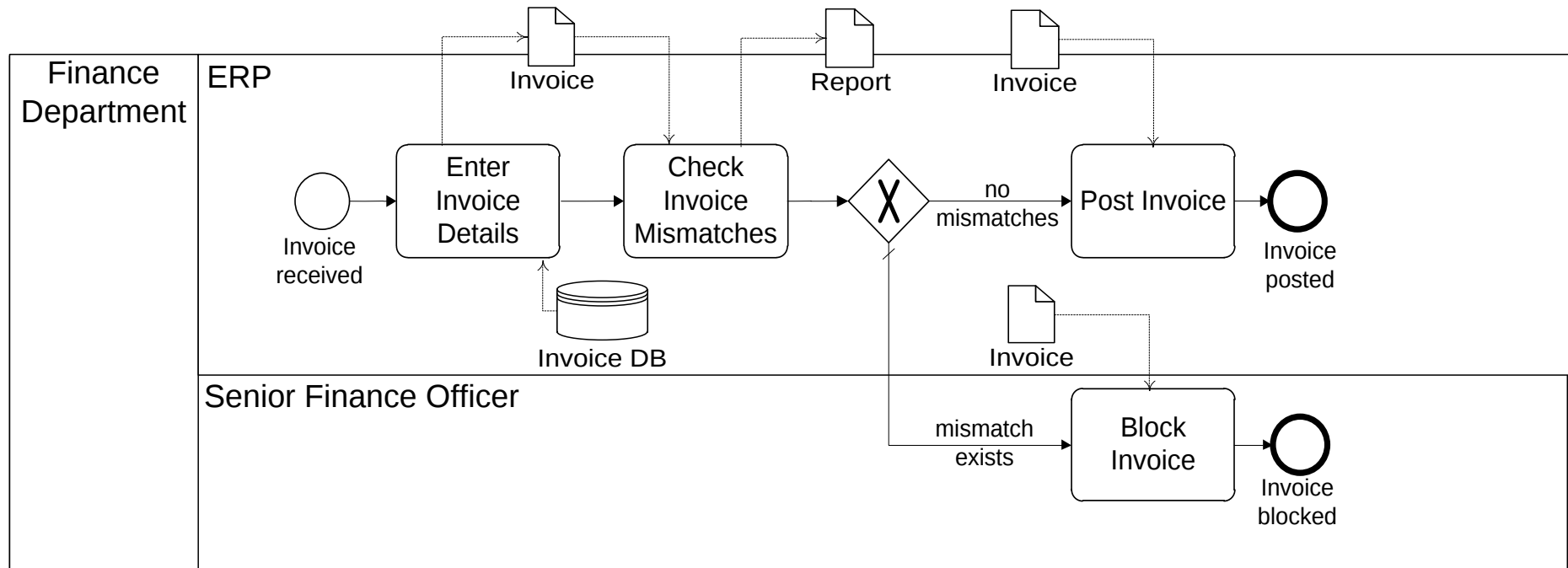
gateway



Flujo de
secuencia

Business Process Model and Notation (BPMN)

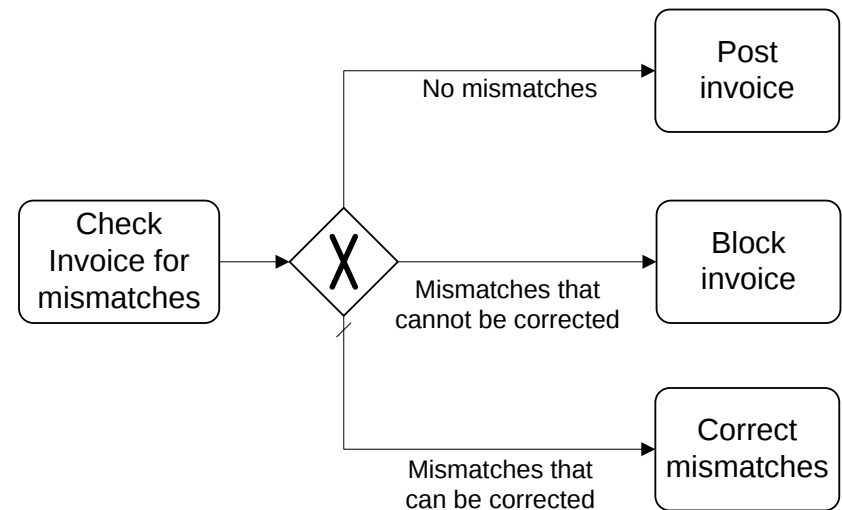
1. ¿Qué se debe hacer y cuándo? - Flujo de control
2. ¿En qué necesitamos trabajar? – Datos
3. ¿Quién está haciendo el trabajo? - Recursos (humanos y no humanos)



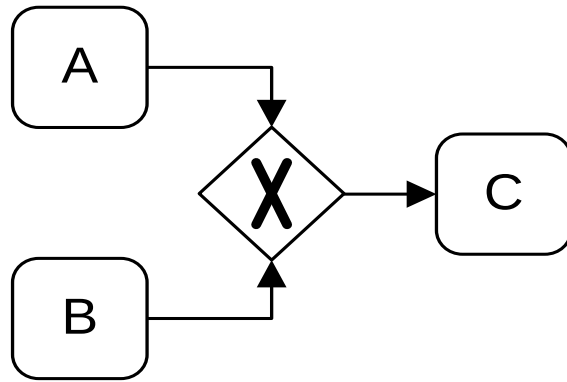
Business Process Model and Notation (BPMN)

Un flujo predeterminado

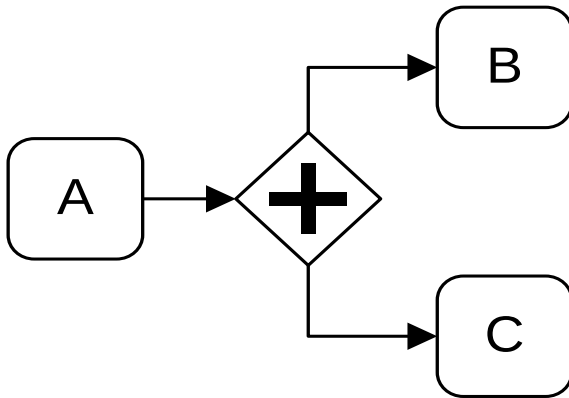
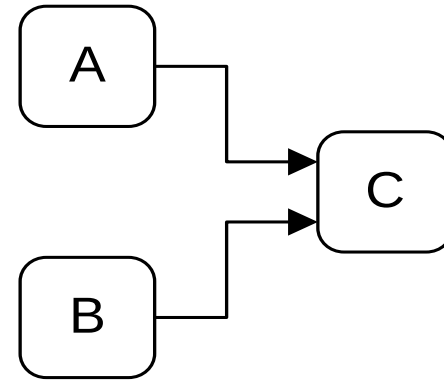
- Un flujo predeterminado es la rama predeterminada que debe elegirse de una división (X)OR-split si todas las otras condiciones se evalúan como falsa
- Evita los interbloqueos si ninguna de las condiciones se evalúa como verdadera debido al error de un modelador
- No es necesariamente la "opción más común"



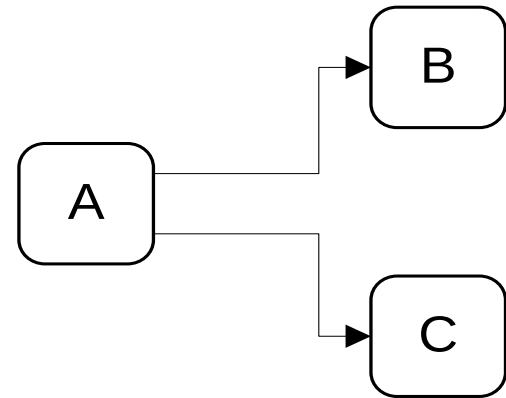
Semántica de flujo de control implícito vs explícito



=

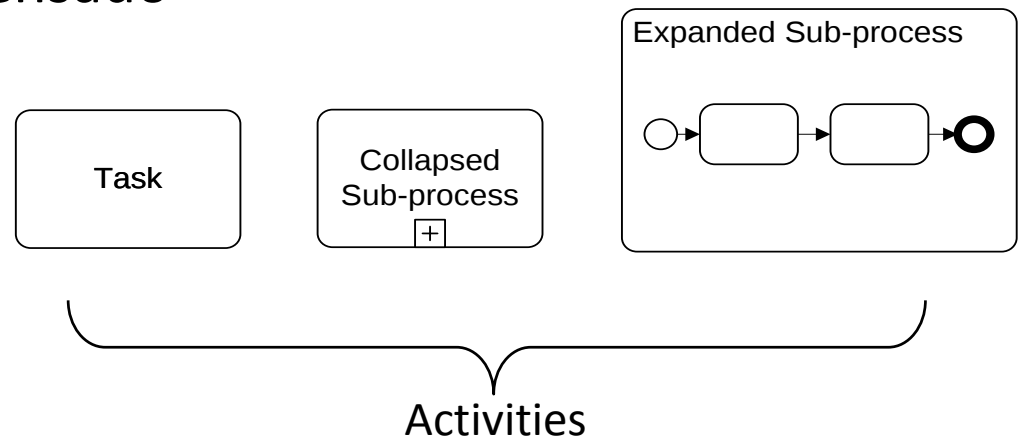


=

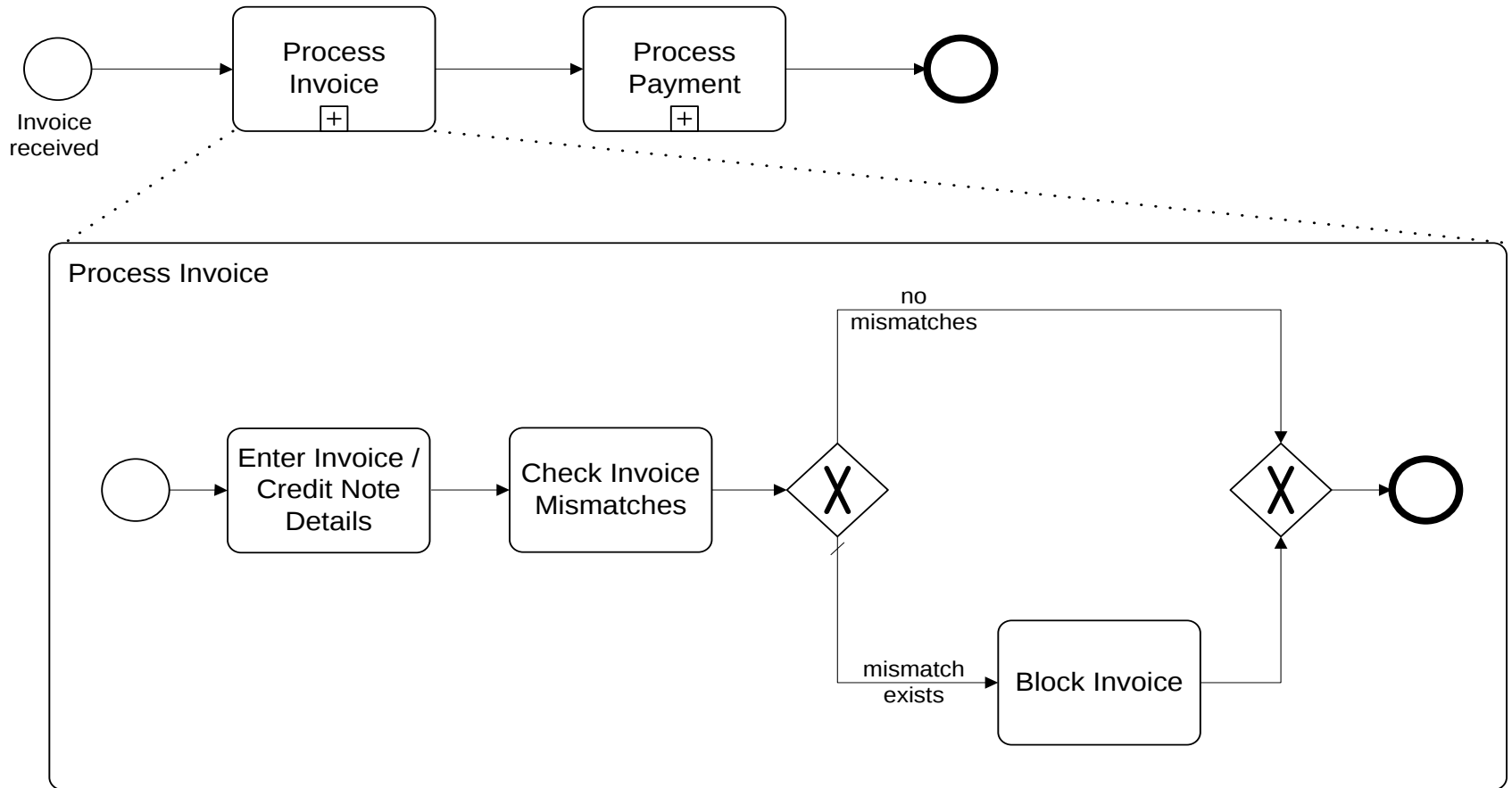


Descomposición del proceso

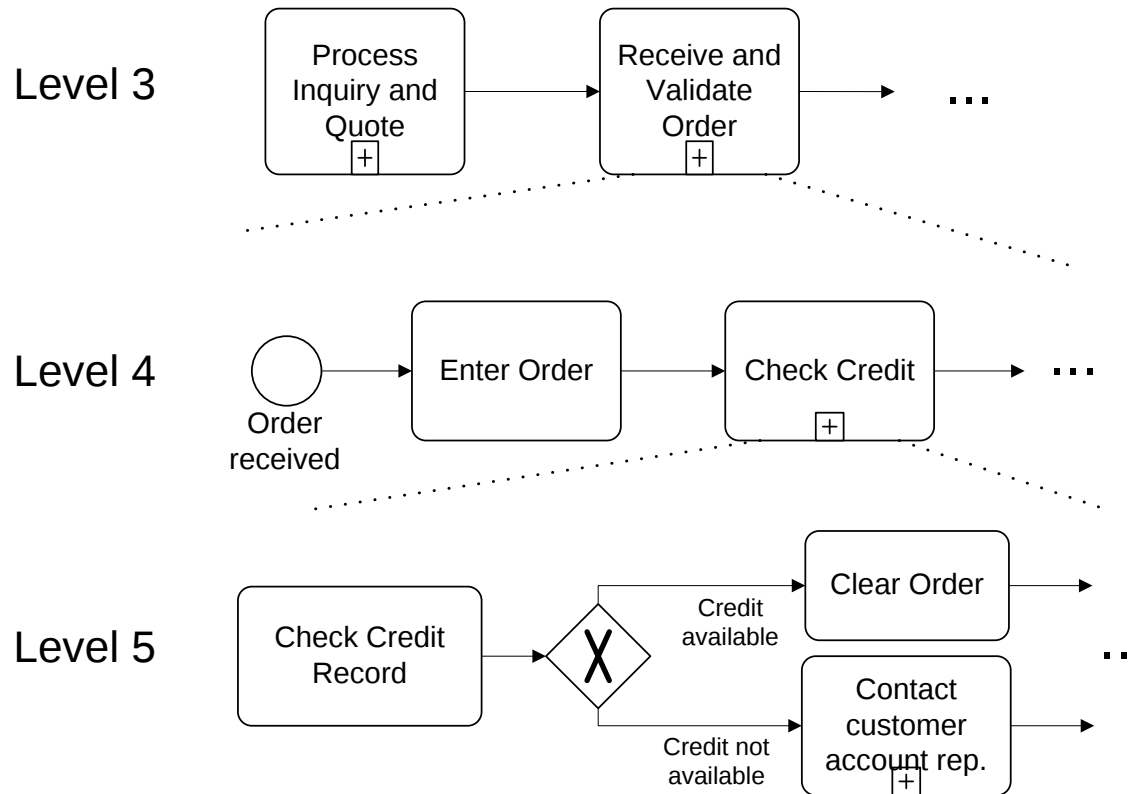
- Una actividad en un proceso puede descomponerse en un "subproceso".
- Usa esta característica para:
 - Dividir los modelos grandes en modelos más pequeños, haciéndolos más fáciles de entender y explicar
 - Identifica partes de un modelo de proceso que deberían ser: Repetido, ejecutado varias veces en paralelo cancelado, o compensado



Ejemplo: Sub-procesos

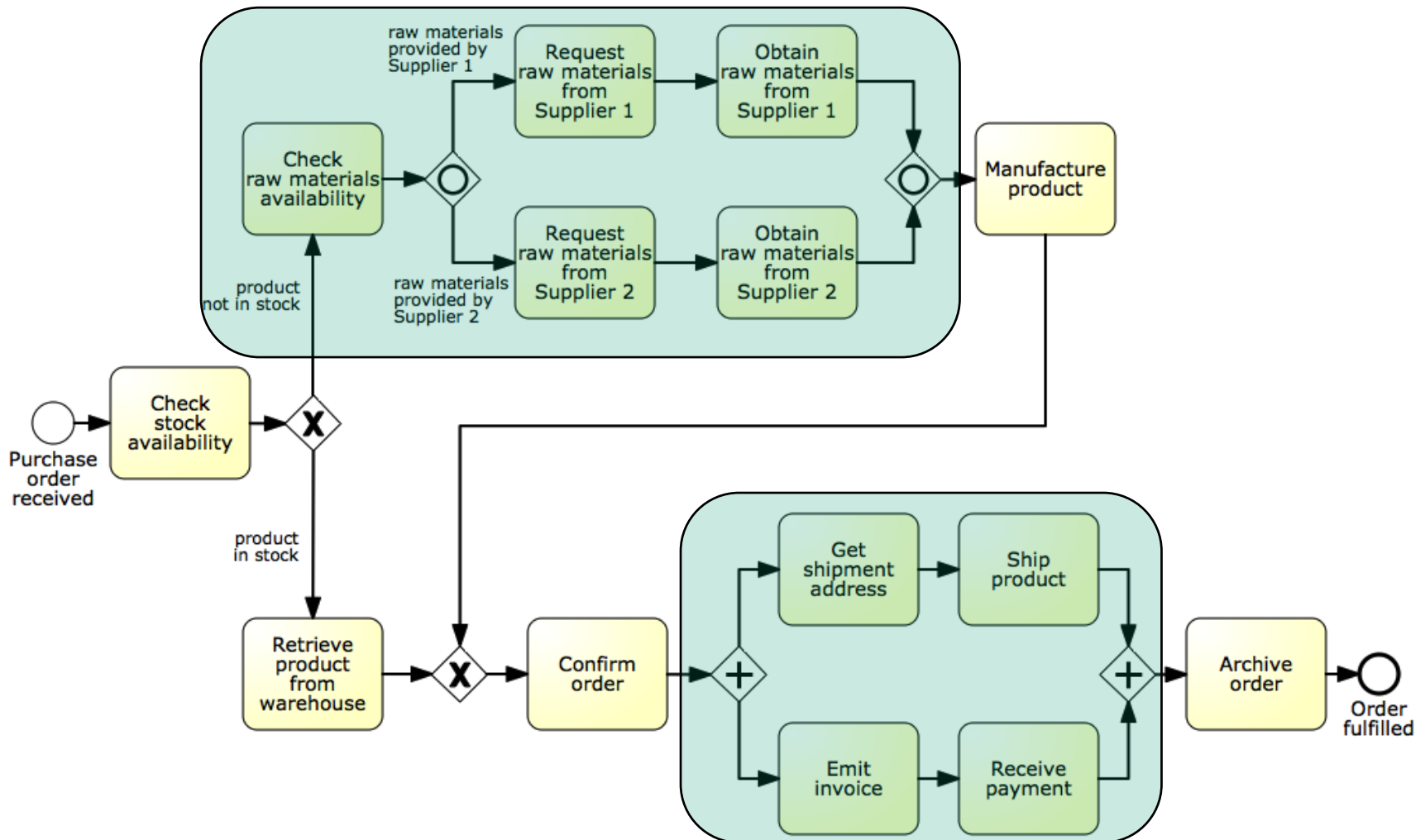


Modelización de jerarquías de proceso con subprocessos



(Fragment of the SCOR reference model)

Identificar posibles subprocessos



Cuestiones de MP

¿Cuándo deberíamos descomponer un modelo de proceso en subprocesos?

Cuando el modelo se vuelve demasiado grande:
Difícil de entender, Aumento de la probabilidad de error

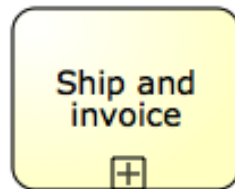
**Regla de oro: no más de 30 objetos de flujo
(actividades, eventos, puertas de enlace)**

Reuso de Procesos

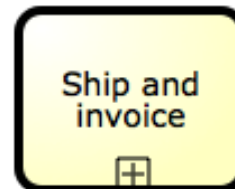
De forma predeterminada, un subprocesso se "incrusta" en su proceso principal

Para maximizar la reutilización, es posible "extraer" el subprocesso y almacenarlo separado en el repositorio de modelo de proceso

Tal subprocesso se llama modelo "general", y se invoca a través de una actividad de "llamada" (representada con un borde más grueso)

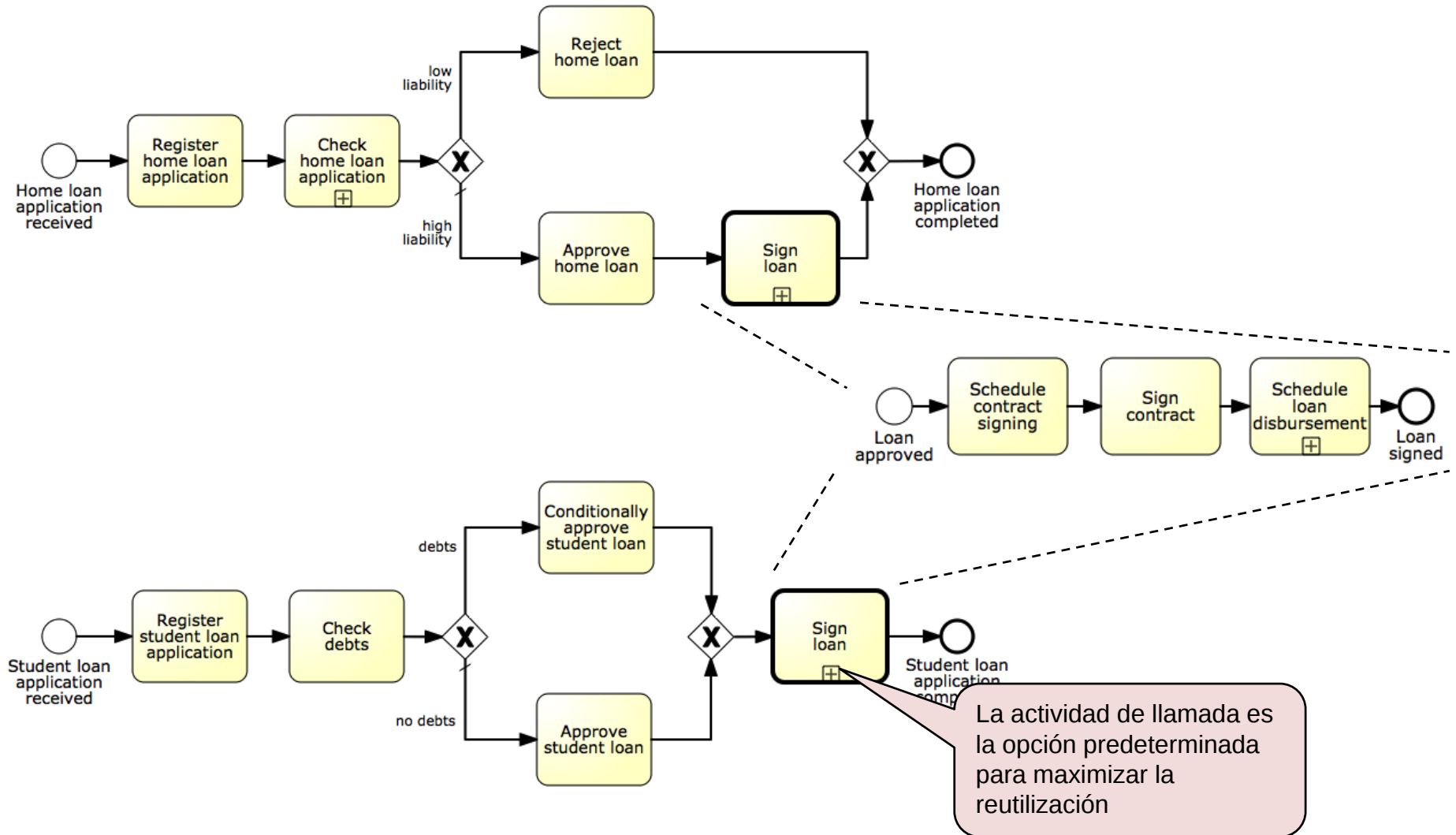


Actividad
(normal)



llamar
actividad

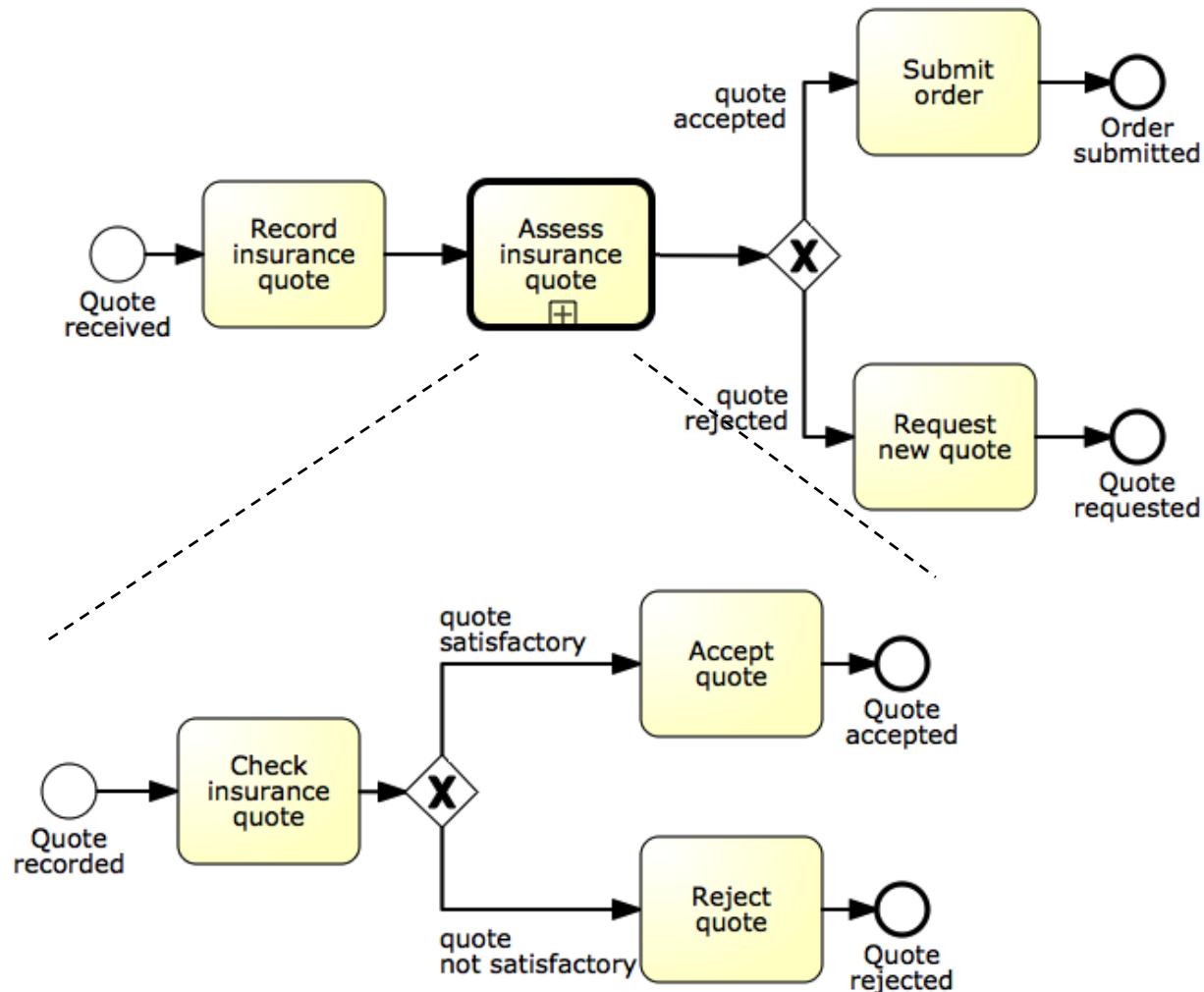
Reuso de Procesos



Reglas sintácticas para subprocesos

- Comience con al menos un evento de inicio
 - Si múltiple, primero ocurre activará el subproceso
- Termine con al menos un evento final
 - El subproceso se completará una vez que todos los tokens hayan alcanzado un evento final. Puede necesitar an (X)OR-split después del subproceso para comprender qué evento (s) final se han alcanzado
- Los flujos de secuencia no pueden cruzar los límites del subproceso
 - Usa eventos de inicio / finalización
- Los flujos de mensajes pueden cruzar los límites del subproceso
 - Para indicar los mensajes que provienen / entran en el subproceso

Ejemplo: subprocesso con múltiples eventos finales

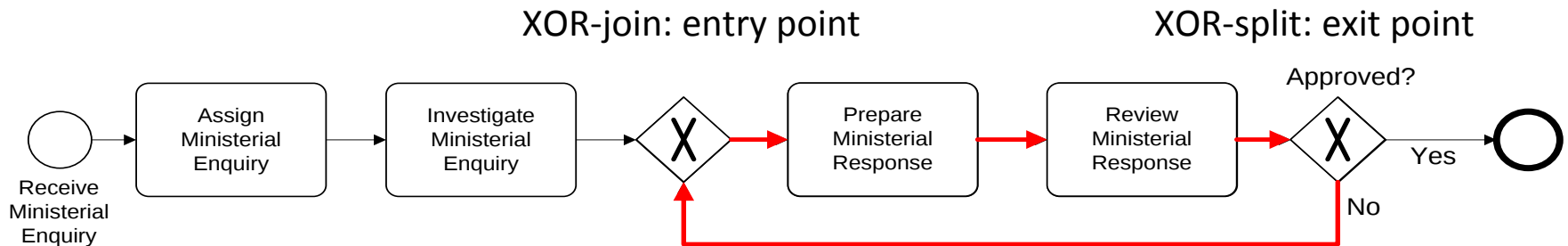


Ciclos

- Un ciclo es un flujo que regresa a un punto "anterior" del proceso.
- Se usa para modelar partes del proceso que se pueden repetir.

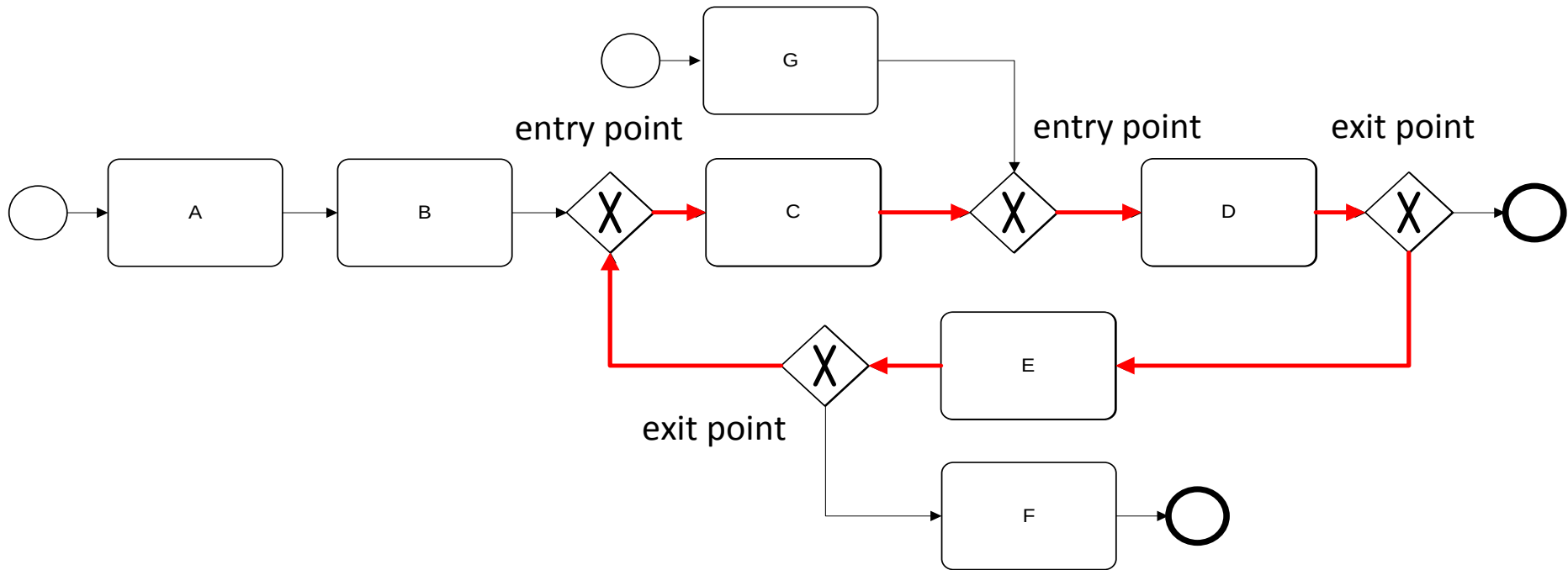
Ejemplo: dirección correspondencia ministerial

El subproceso Finalizar la respuesta ministerial incluye la preparación de la respuesta ministerial y la revisión de la respuesta por parte del registrador principal. Si el Registrador no aprueba la Respuesta, esta última debe prepararse nuevamente para su revisión. El proceso finaliza solo una vez que la Respuesta ha sido aprobada.



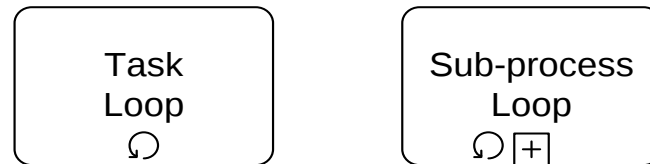
Ciclos

- Arbitrario = no estructurado, es decir, puede tener múltiples puntos de entrada y salida



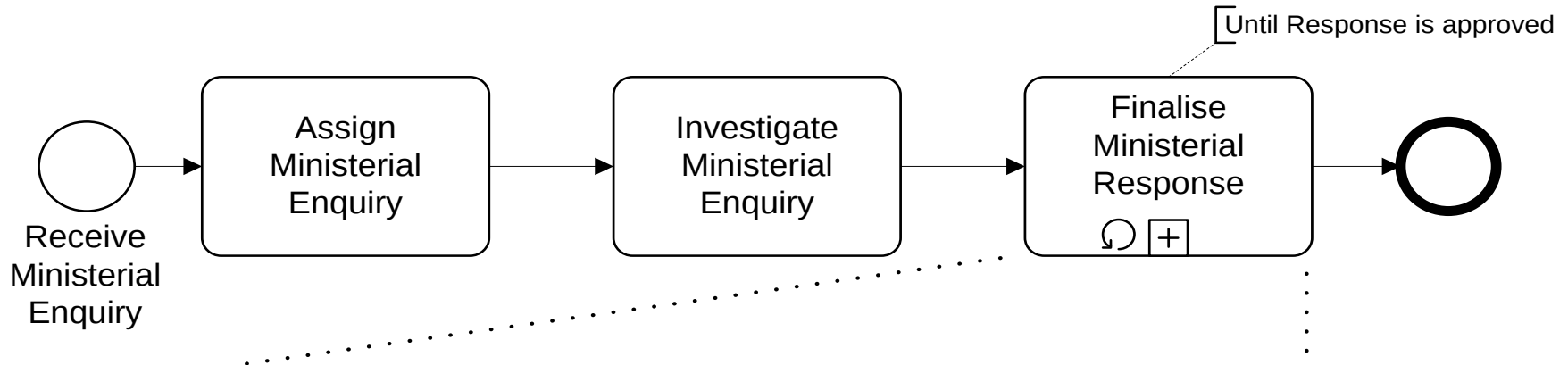
Repetición estructurada en bloque: bucle de actividad

- BPMN proporciona el constructo Activity Loop, que permite la repetición de una tarea o un subproceso



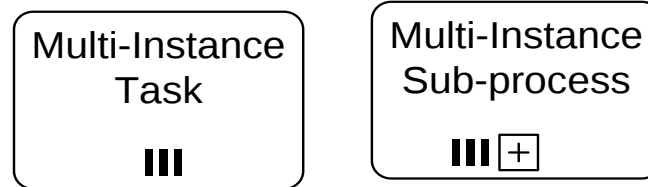
- La principal diferencia es que Activity Loop está estructurado, mientras que los ciclos arbitrarios pueden estar desestructurados
- Sugerencia: use Activity Loop cuando su repetición esté estructurada

Ejemplo Repetición estructurada



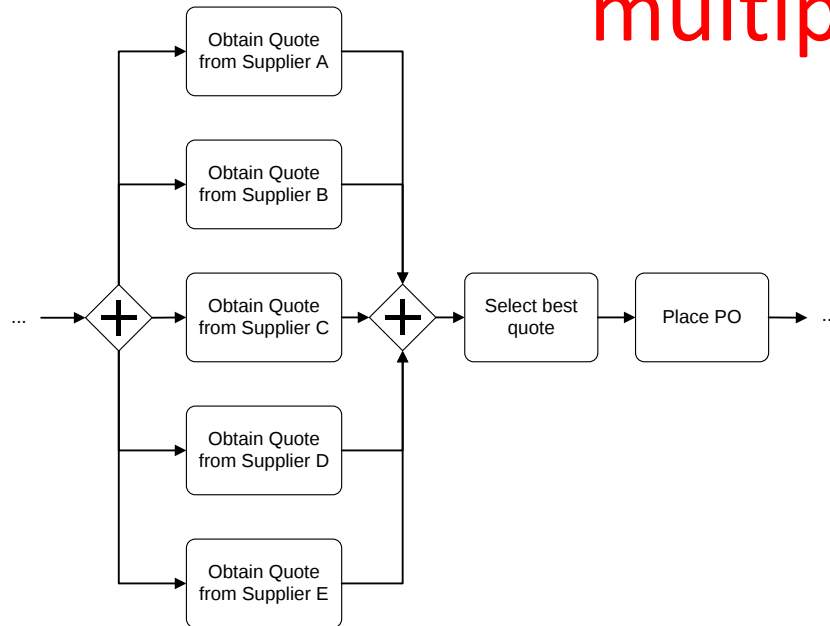
Repetición paralela: actividad de múltiples instancias

- La actividad de varias instancias proporciona un mecanismo para indicar que una actividad se ejecuta varias veces al mismo tiempo

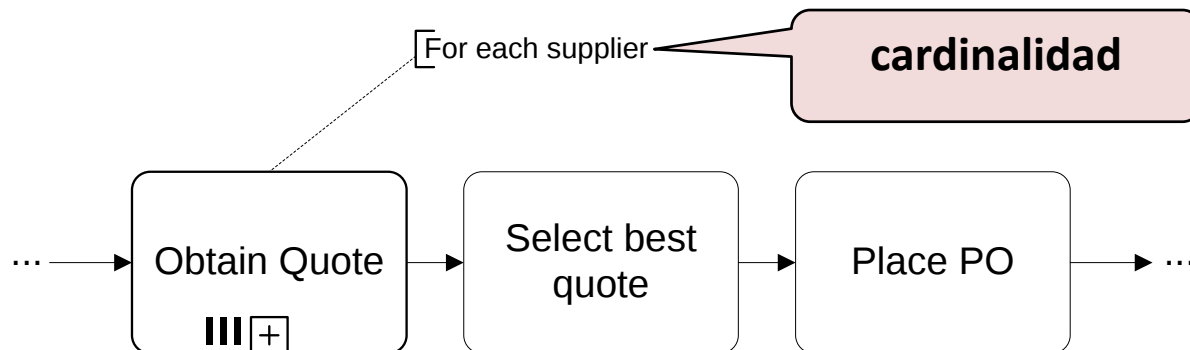


- Útil cuando la misma actividad necesita ser ejecutada para múltiples entidades o elementos de datos, tales como:
 - Solicitar cotizaciones de múltiples proveedores
 - Comprobar la disponibilidad de cada elemento de una línea de producción
 - Enviar y recopilar cuestionarios

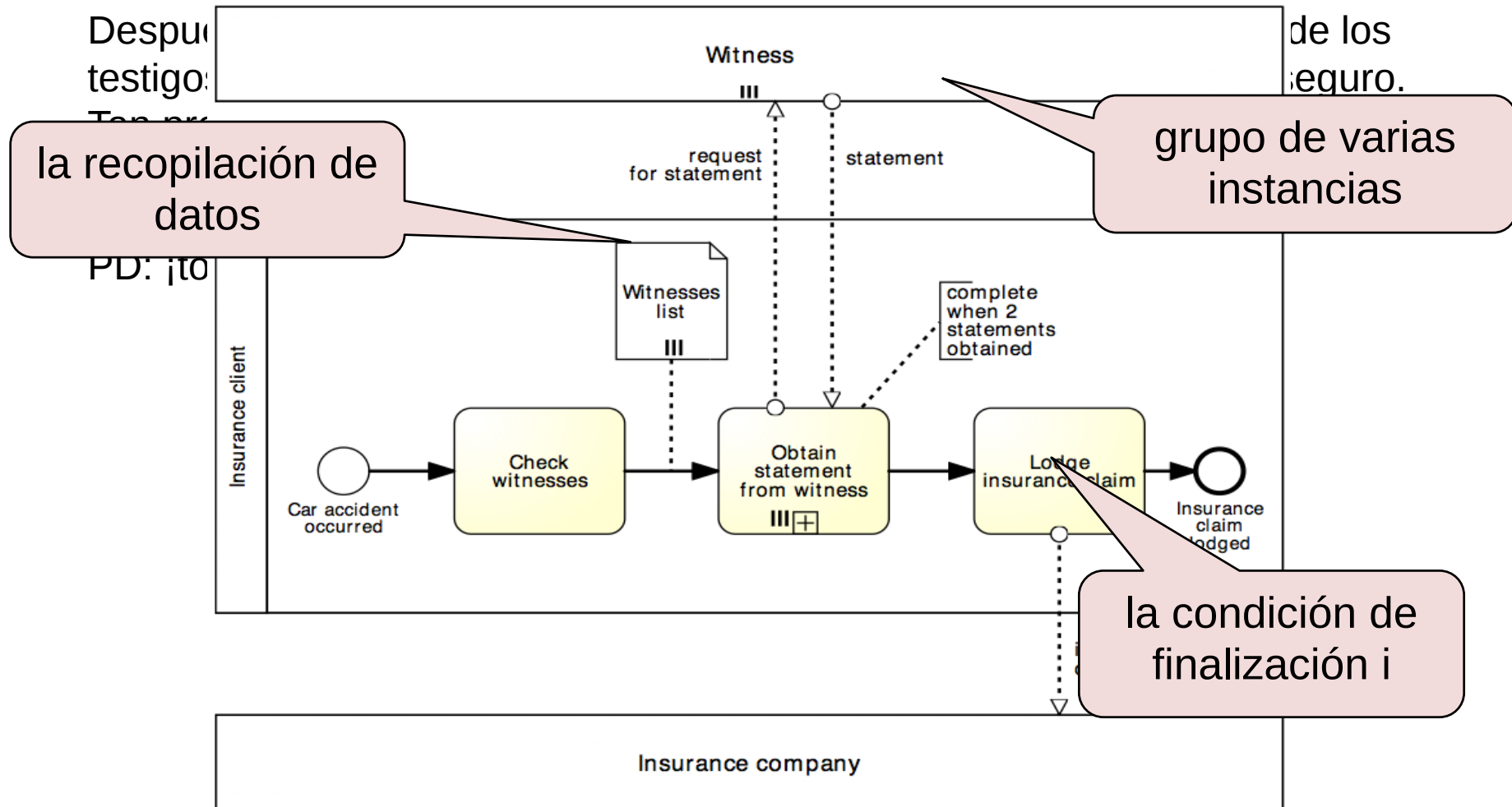
Solución: sin actividad de instancias múltiples



Solución: sin actividad de instancias múltiples



Ejemplo: actividad multi-instanciada



Minería Semántica

Jose Aguilar

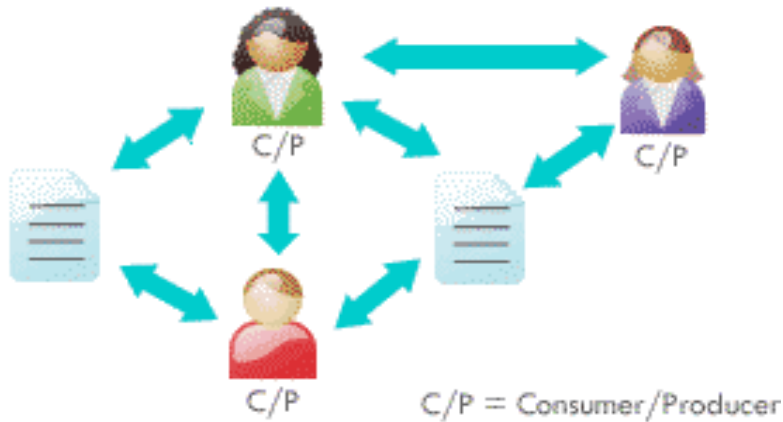
WEB *.0 -> WEB 3.0

Web 1.0



- Static pages
- HTML.

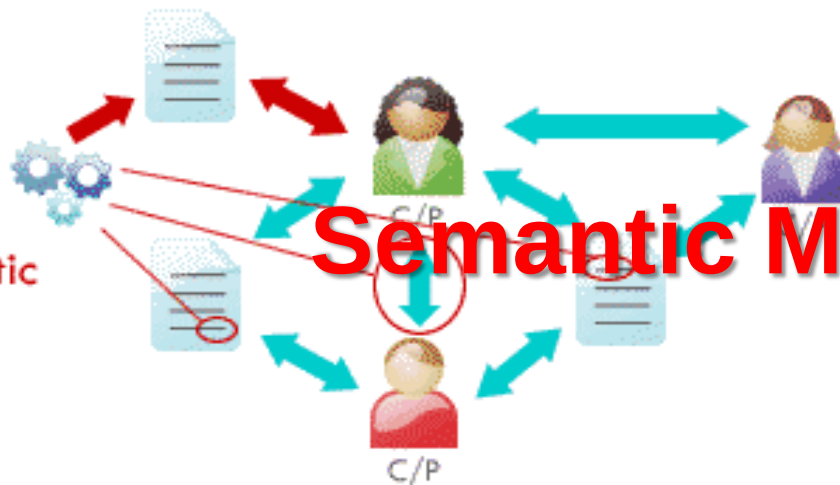
Web 2.0



3 basic principles

- The web as a platform
- Take advantage of Collective Intelligence
- Enriching experiences of the user

The Semantic Web

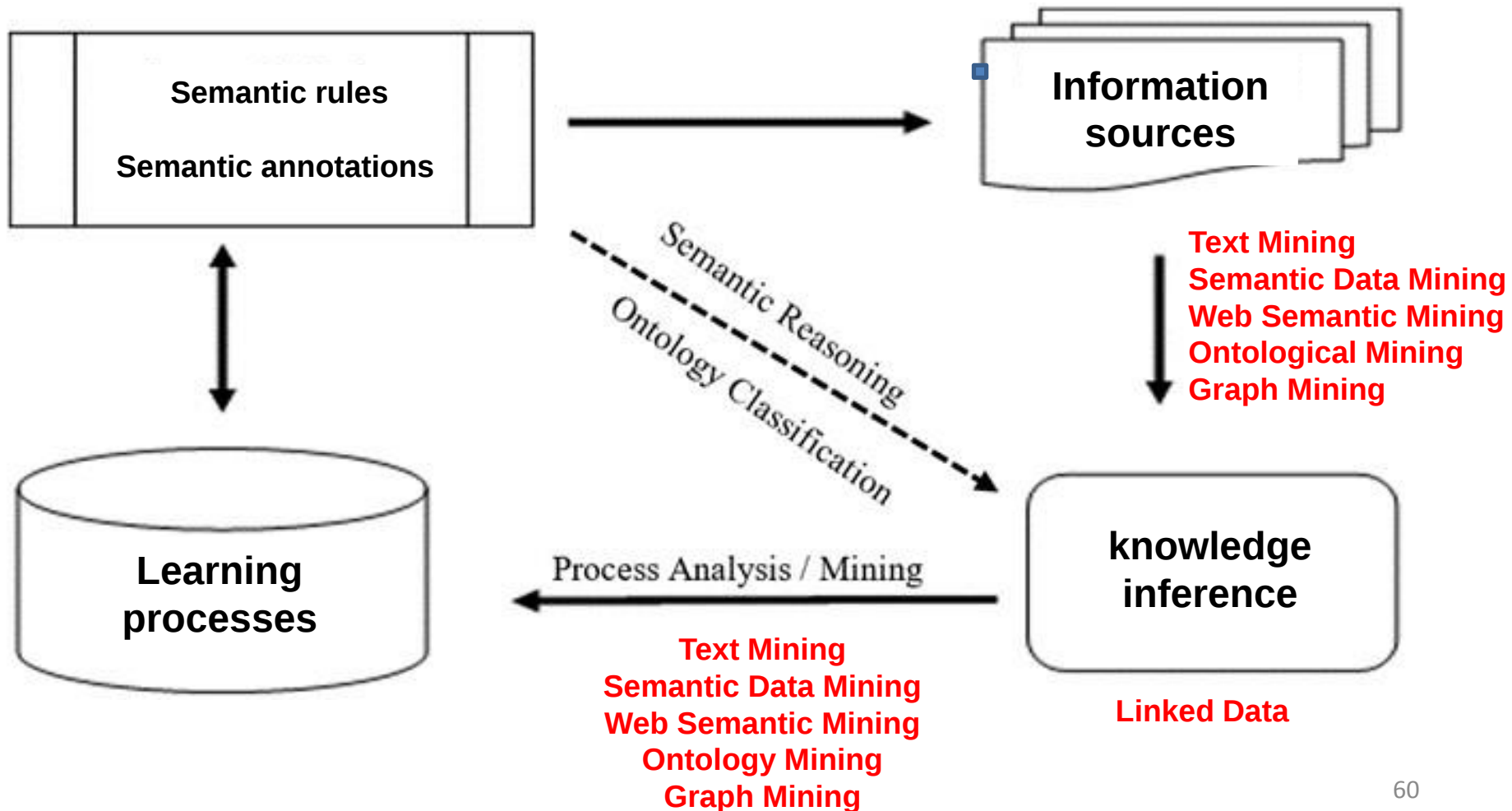


Semantic Mining

Web 3.0 is based on

- A more "intelligent" Internet
- Users carry out searches close to natural the
- Information has associated semantics
- The website deduces information through rules associated with the meaning of the content

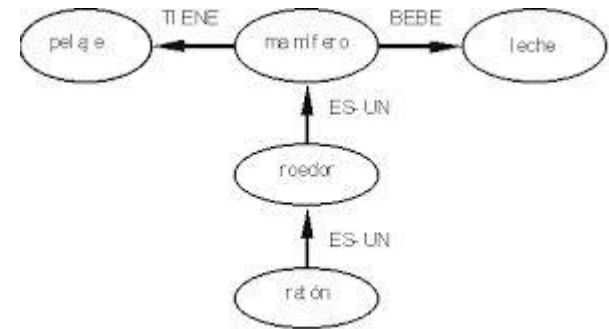
General idea

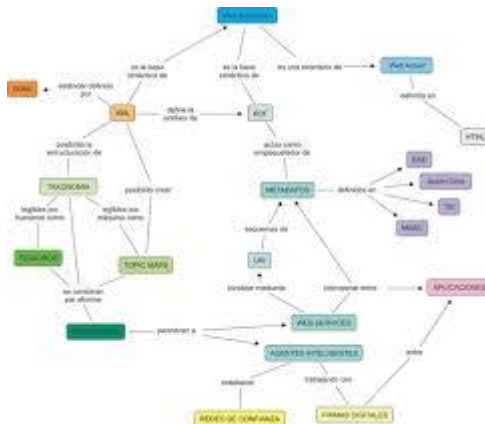
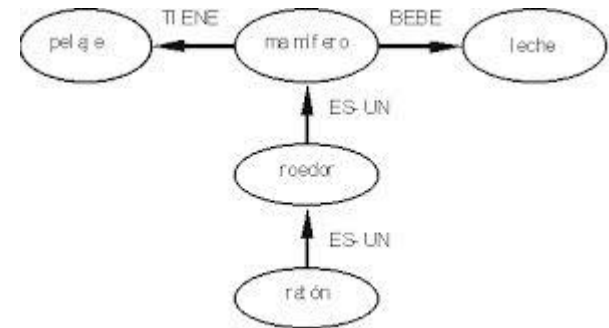


Semantic Mining



	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	\$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	\$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	\$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874235	BOSA	\$ 130.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	\$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	\$ 800.000
8	Camila Davalos	Cajera	3214675961-7584621	TEUSQUILLO	\$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	\$ 600.000
10	Pamela Carrasco	Mesera	3157485912-2485796	CANDELARIA	\$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	\$ 600.000
12	Jairo Arevalo	Parquedero	3002157459-2861459	BOSA	\$ 489.500
13				TOTAL	\$ 8.199.500





- 62

Web Semantic Mining

Web Semantic Mining

It is the integration of two areas of knowledge:

- **Semantic Web**
- **Web Mining**

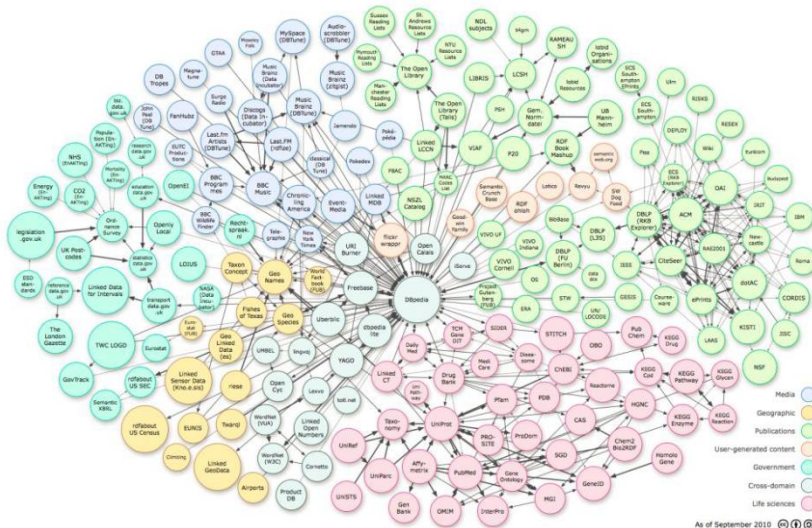
The **Semantic Web** is used to give meaning to the data found on the Web.

Web Mining is used to extract patterns of behavior on the Web.

Web Semantic Mining

Change of paradigm from data mining to knowledge mining

Semantic Web Mining: **Mining of knowledge in the web** encoded in domain ontologies, etc.



Types of semantic resources

- Domain Ontologies
- Ontologies in the semantic web

Web Mining

There are several types :

- The content of the web
- The structure of the web
- The use made of the web.



Search results
Content of the website

Text Mining

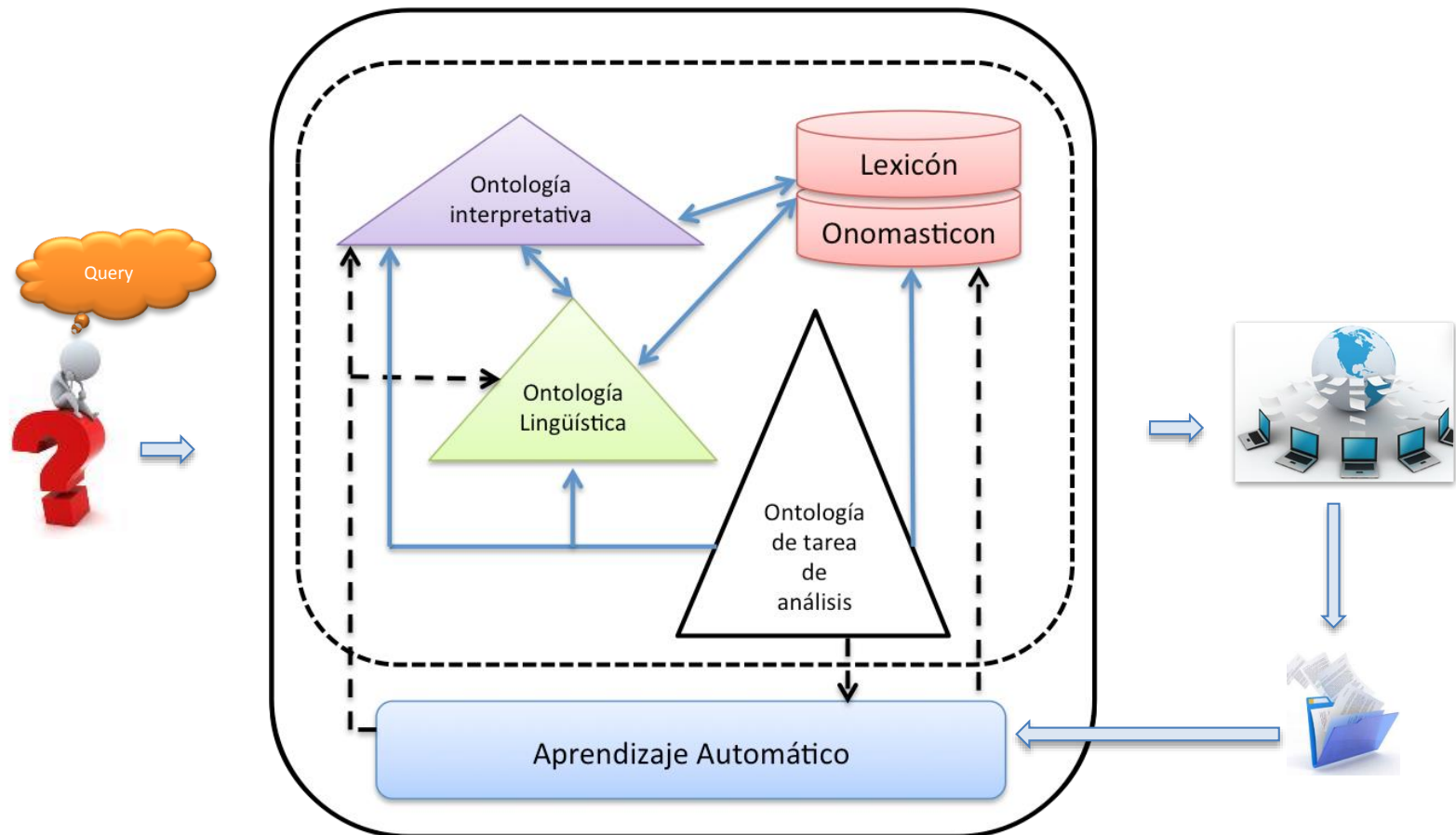
Links

Graph Mining

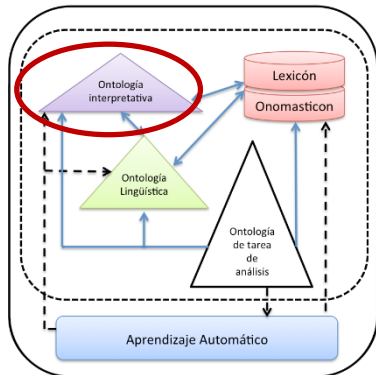
General patterns of use
Personal access patterns

User Mining

Dynamic Semantic Ontological Framework

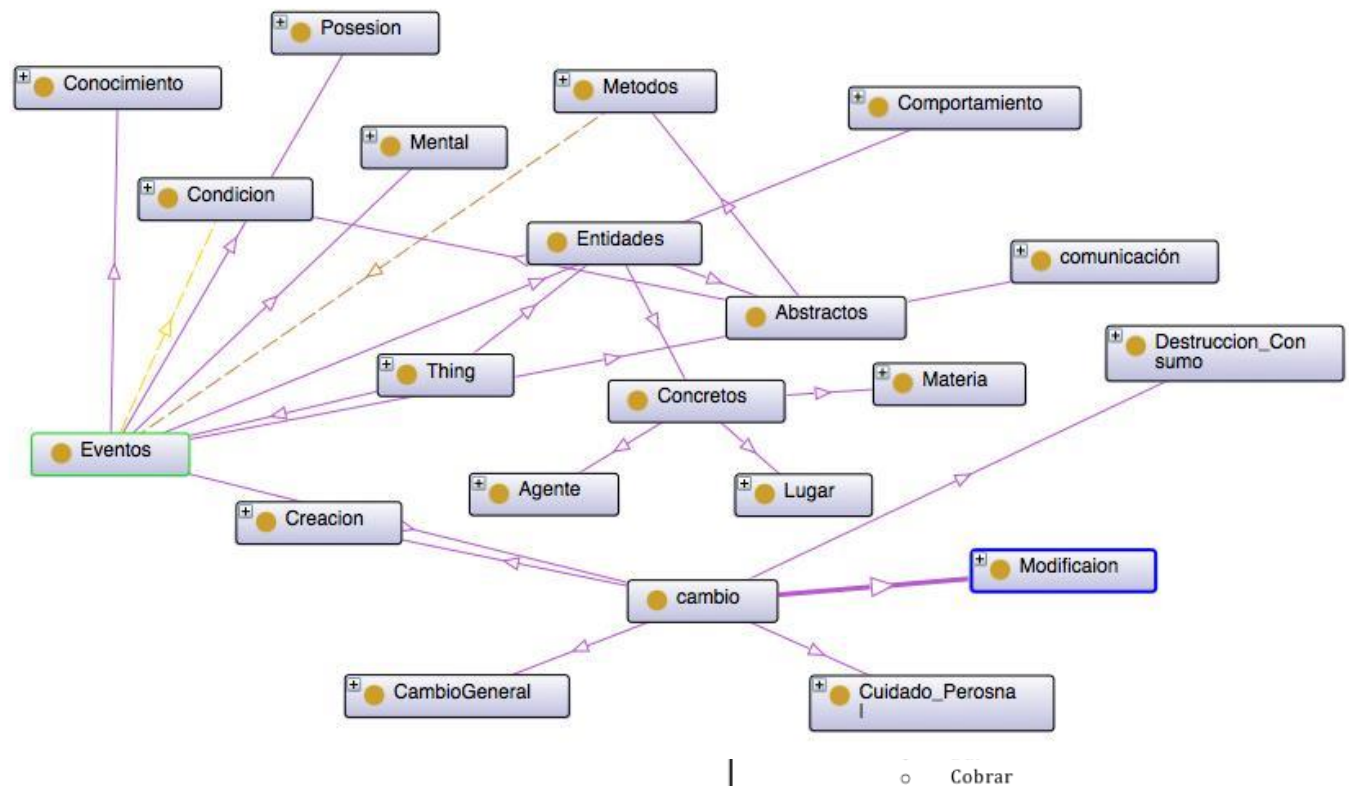


Interpretive Ontology

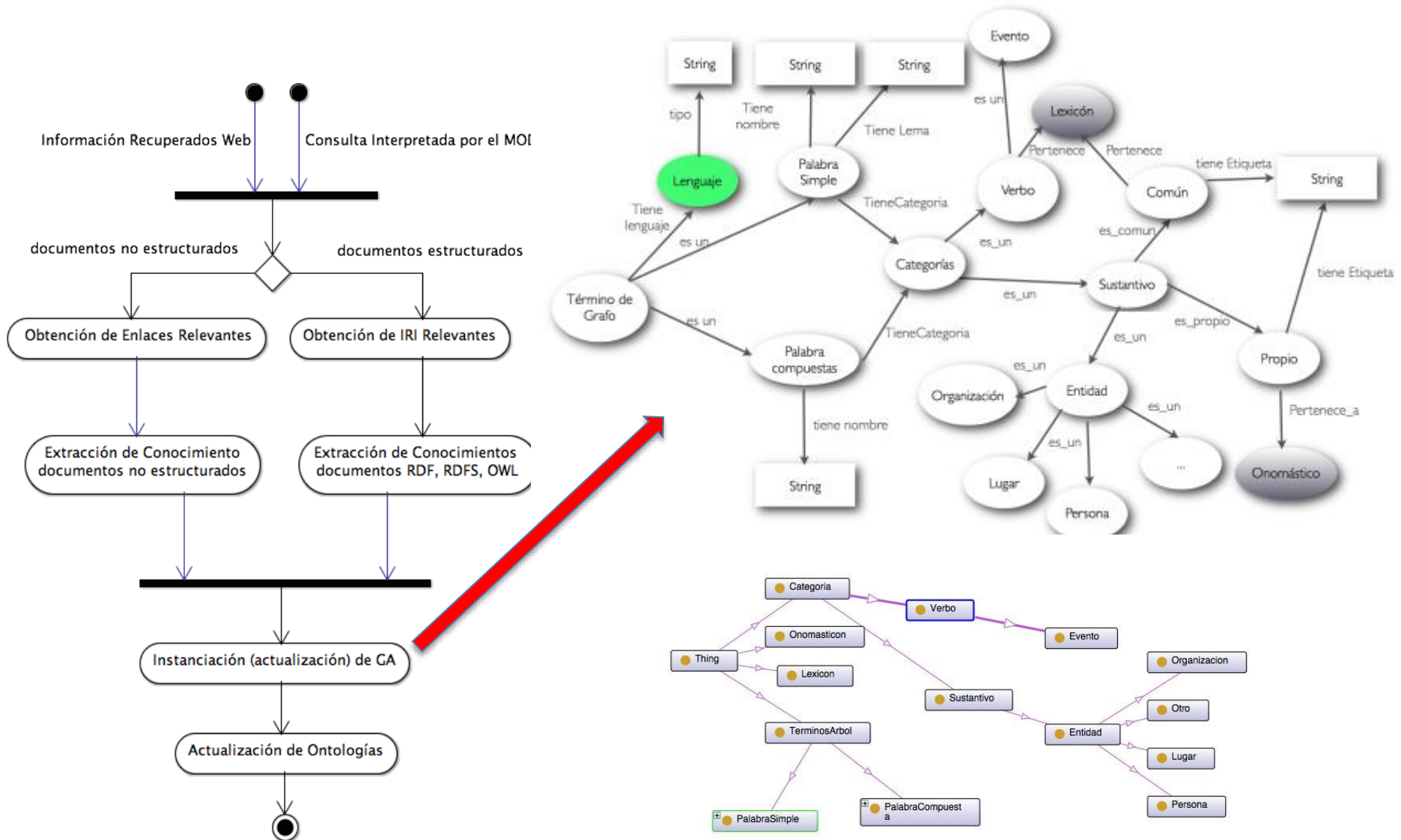


MODS ENTIDADES	Definición [10]
Abstractos	Los abstractos pueden ser números, conjuntos, definiciones
Concretos	Los concretos son objetos físicos o que se pueden definir en algo específico (por ejemplo. el planeta Venus. ese árbol.

MODS EVENTOS
Comunicación
• General (Decir, hablar) • Valoración (Criticar, felicitar) • Mandato (Suplicar, ordenar)
Comportamiento

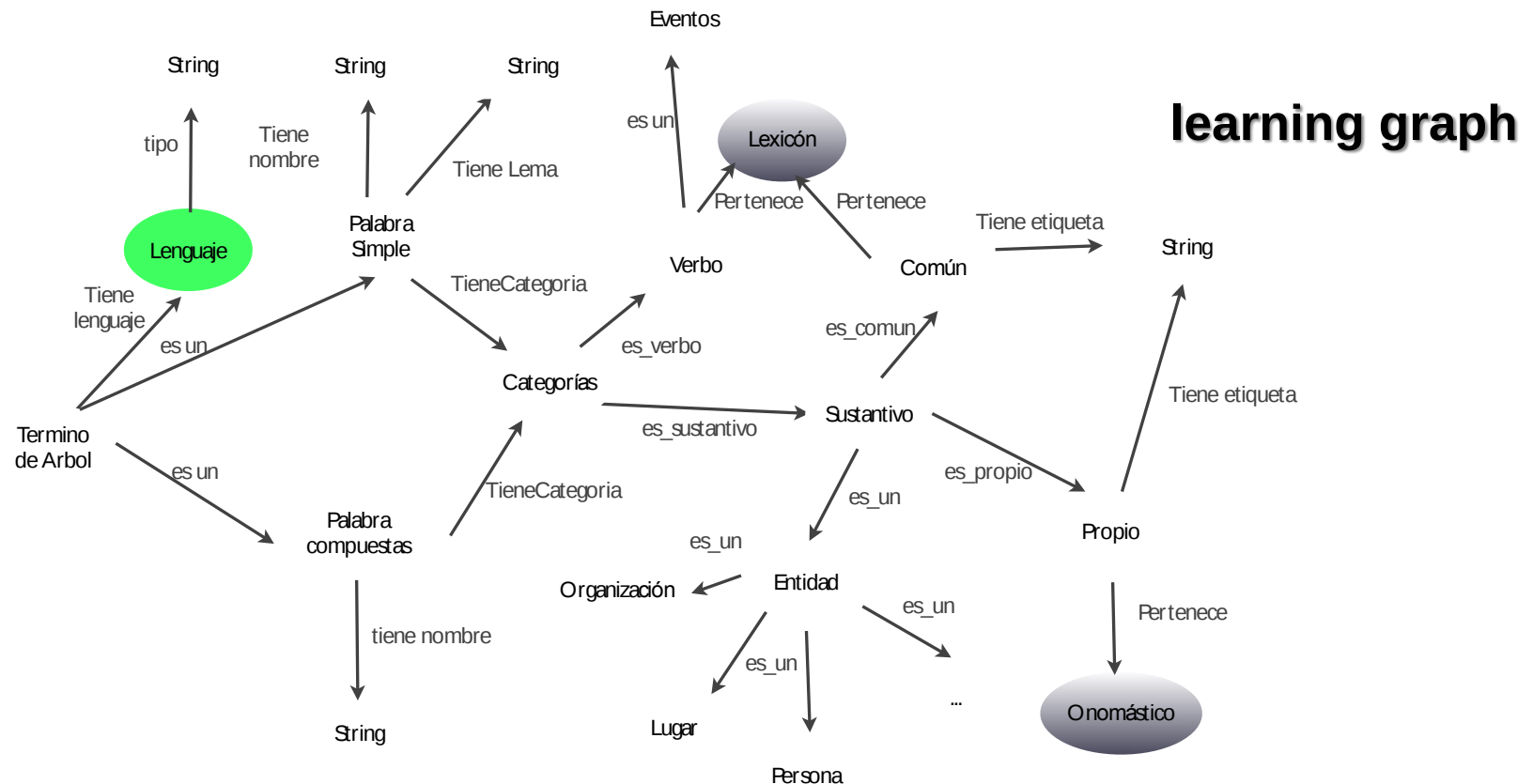


Learning Graph

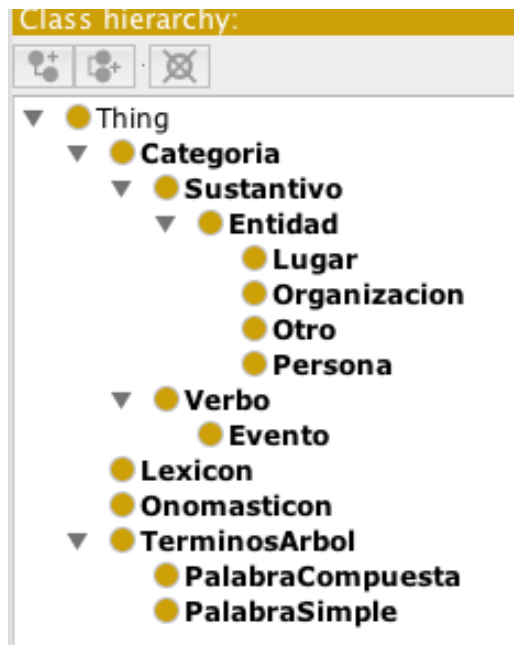


Semantic Mining based on the Learning Graph

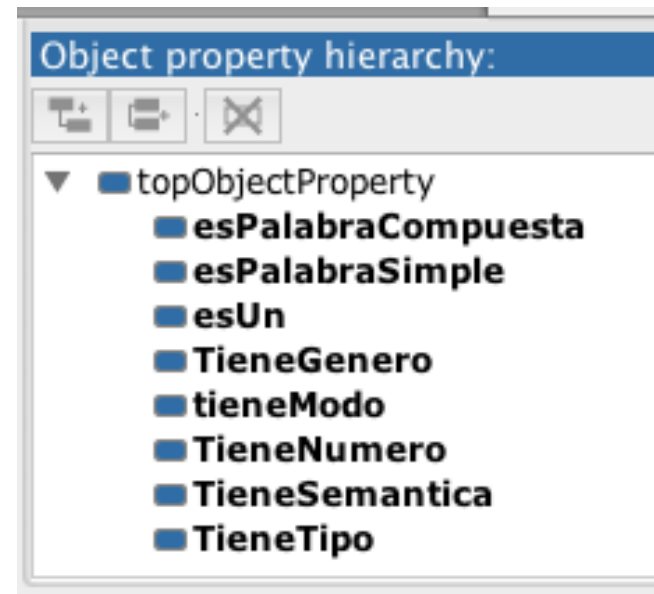
The learning graph is composed of a set of basic axioms to infer new knowledge.



Semantic Mining based on the Learning Graph



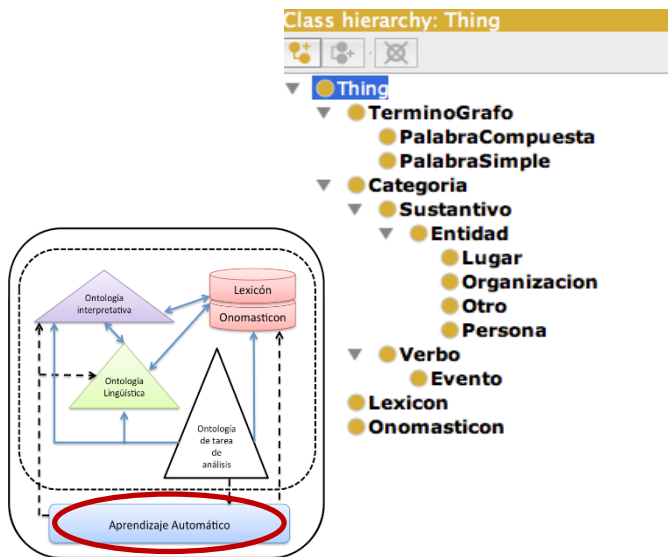
The classes defined in the learning graph



The relationships defined in the learning graph

Learning Graph

The graph has **the entities and relationships relevant**, which is calculated



$$w_j = tf_j * fi_j$$

tf_i average TF-IDF in the processed texts

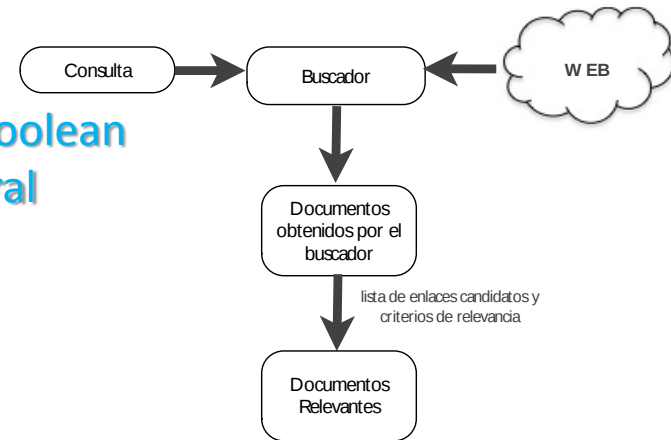
$$fi_j = \log\left(\frac{\text{processed texts}}{\text{texts with the term}}\right)$$

It will be relevant if its weight is greater than or equal to the average of the weights.

Query with MODS: Extraction of Relevant Documents from the Web

Relevant documents according to the degree of similarity between the user's query and the recovered documents

To determine this similarity, different models can be used : **Boolean expressions, a vector model, those based on fuzzy logic, neural networks or Bayesian networks, etc.**



Vector model: This model represents the query and the documents as vectors. Thus, a vocabulary of size t will define a t -dimensional space such that:

- a document d_j is represented by a vector
- a query q is represented as a vector

$$d_j = (w_{1j}, \dots, w_{tj})$$

$$q = (w_{1q}, \dots, w_{tq})$$

Query with MODS: Extraction of Relevant Documents from the Web

Weighting of the terms of the documents and of the query

Frequency analysis, number of occurrences of the terms found in the query and in the recovered documents,

TF-IDF weights, the importance of a term to discriminate the document and/or collection of documents.

$$IDF(term) = \log(N/DF)$$

$$TF - IDF = TF(term) * IDF(term)$$

Where: N = number of documents in the collection, DF = number of documents in which the term appears, TF = Frequency of appearance of the term in the document

Frequency and weights determine the similarity

Query with MODS:

Extraction of Relevant Documents from the Web

Suppose a user performs the following query in natural language in the google search engine:

“Universidad de Los Andes de Mérida”

MODS gets the following Boolean query:

(“Universidad de los Andes” and Mérida and Venezuela) or (ULA Mérida and Venezuela) or (“Universidad de los Andes” and “Núcleo Mérida” and Mérida and Venezuela) or (ULA and “Núcleo Mérida” and Mérida and Venezuela),

This query is made in the Google search engine, and MODS gets the set of links to documents

[http://es.wikipedia.org/wiki/Universidad_de_Los_Andes_\(Venezuela\)](http://es.wikipedia.org/wiki/Universidad_de_Los_Andes_(Venezuela))
<https://www.facebook.com/ula.venezuela>
<https://www.facebook.com/pages/Facultad-de-Ingenier%C3%ADa-ULA-Venezuela-Sitio-Oficial/258084854230578>
http://www2.ula.ve/andes/images/stories/inf_gestion_cap_i.pdf
<http://lama.adm.ula.ve/pderecho/>
http://www2.ula.ve/andes/images/stories/pdula_capi.pdf
<http://www.venezuelaaia.com/2014/02/decanos-de-ula-merida-estudian-suspender-clases/>
<http://www.venezuelaaia.com/laula/>
<http://www.slideshare.net/MANUELLITOR>
<http://www.noticias24.com/Venezuela/noticia/221873/suspender-las-actividades-academicas-indefiniidamente-en-la-universidad-de-los-andes/>
http://lar.ask.com/web?z=ula+de+merida+venezuela&qsrc=999&sem&siteid=1332&enc=utf_8&fr=1&ad=sem&an=google_s&my=b&kwd=ula%20de%20merida%20venezuela&res&res=3393039999&ia=8&mob=8&sc=8&aid=8&op=211&kwid=40510943168&agid=9114244999&date=20131204
<http://www.mcti.gob.ve/Noticias/16134>
<http://www.actualidadyente.com/noticias-de-merida-venezuela/32-academicas/12888-hoy-la-ula-cumple-229-anos-de-fundada>
<http://www.actualidadyente.com/noticias-de-merida-venezuela/48-informacion-general/merida/12146-ula-crea-observatorio-de-derechos-humanos>
<http://www.aluniversa.com/nacional-y-politica/140210/decanos-de-ula-merida-podran-suspender-clases-por-violencia>
<http://eluniversitario.net/ula-suspenden-actividades-academicas-indefiniidamente-y-administrativas-hasta-el-lunes-17-de-febrero-en-el-nucleo-de-merida/>
<http://canalnoticia.com.ve/index.php/noticias-venezuela/item/26049-decanos-de-ula-merida-podran-suspender-clases-por-violencia>
<http://carlosramosnivas.com/2013/07/24/tesoros-de-merida-universidad-de-los-andes-primera-universidad-republicana-de-latinoamerica/>
http://www.acdes.ula.ve/spa0.php?doc=S1316-49162007000400019&script=scr_artes
<http://www.redalc.org/pdf/775/70504709.pdf>
<http://pod.ula.ve/portal/>
<http://www.almomento360.com/portal/estudiantes-de-la-ula-merida-protestan-en-contra-de-la-inseguridad/>
http://www.linkedin.com/search/?orig=TSEO_S&firstName=Judi&lastName=VeigaM_Gover%3A0&tr=TSEO_S&H
<http://meridemusical.wordpress.com/actas-aguacalones/voces-de-santa-rosa/>
<http://vevuanoticias.com/universidad-de-los-andes-suspende-actividades/>
<http://www.minci.gob.ve/2012/11/entregan-financiamiento-a-investigadores-peil-de-merida/>
<http://noticiasvenezuela.info/2013/12/01/ula-sarueba-maestria-en-ciencias-de-la-actividad-fisica-y-los-deportes-para-la-ula-merida/>
<http://www.laexpresica.com/51434-venezuela-rector-ula-rechaza-ingreso-grupos-armados-mantiene-suspension-clases>
<http://informe21.com/universidad-los-andes>
<http://www.iberamerica.net/venezuela/prensa-generalista/noticias24.com/20140213/noticia.html?id=QG0812>
http://maribelisvarsmarcha.blogspot.com/2014_02_09_archive.html
<http://reimendenzaco.com.ve/mediante-encuesta-evaluaron-la-situacion-para-un-possible-reinicio-de-actividades-en-la-ula/>

Query with MODS: Extraction of Relevant Documents from the Web

Accuracy: proportion of documents retrieved that are relevant.

$$\text{Accuracy} = \frac{\text{Recovered relevant documents}}{\text{Recovered documents}}$$

	Yahoo	Google
Accuracy	0,2716	0,3699

Our system gives all documents as relevant
(100% accuracy).

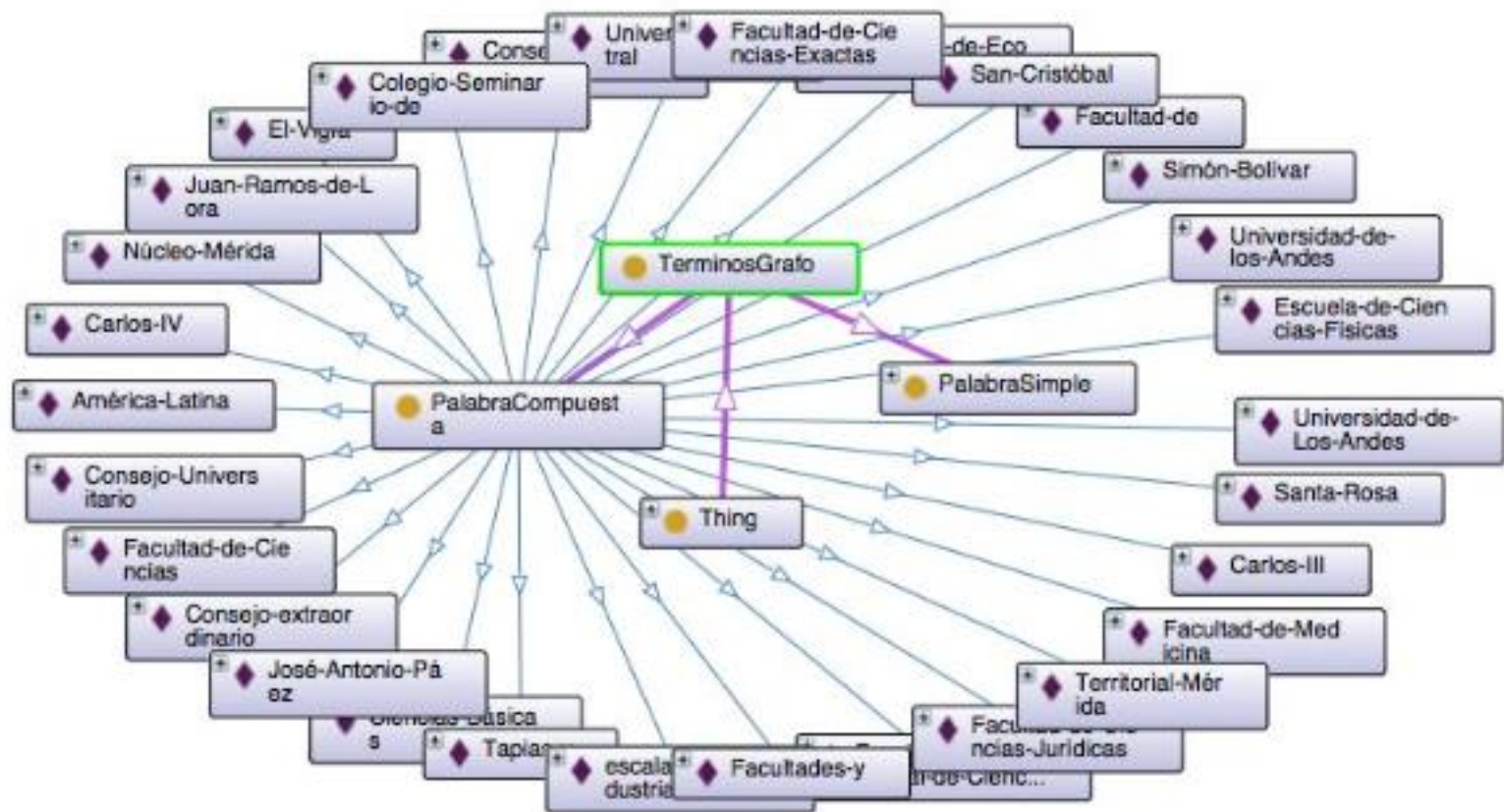
Random Queries	Yahoo	Google
Query 1	0,2456	0,311
Query 2	0,299	0,323
Query 3	0,253	0,312
Query 4	0,2612	0,3419

Web Mining based on the Learning Graph

- Automatic creation of the **terminological base** of a domain
- **is_a** Relationship detection (superclass-subclass),
- Creation of **electronic lexicons** for nouns and verbs.

Semantic mining

- ✓ Build a **Terminological Base** in a specialized domain.



Semantic Mining based on the Learning Graph

Description: LexiconElectronico	
Members +	
◆ ADMISIÓN	?
◆ APROBADO	?
◆ ARTÍCULO	?
◆ ASPIRANTE	?
◆ AÑOS	?
◆ CASO	?
◆ CO-TUTOR	?
◆ COMISIÓN	?
◆ CONOCIMIENTO	?
◆ CRITERIOS	?
◆ CRÉDITOS	?
◆ CUMPLIR	?
◆ DEBE	?
◆ DEBEN	?
◆ DEBERÁ	?
◆ DEBERÁN	?
◆ DESARROLLAR	?
◆ DOCTOR	?
◆ DOCTORADO	?
◆ ELABORAR	?
◆ ES	?
◆ ESTAR	?
◆ ESTARÁ	?
◆ ESTÁN	?
◆ FORMAR	?
◆ GRADO	?
◆ HA	?
◆ HABER	?

Description: LexiconElectronico	
◆ INVESTIGACIÓN	?
◆ JURADO	?
◆ LAPSO	?
◆ MIEMBROS	?
◆ NIVEL	?
◆ NOMBRAR	?
◆ PARTE	?
◆ PAÍS	?
◆ PODRÁ	?
◆ PODRÁN	?
◆ POSEER	?
◆ PRESENTAR	?
◆ PROFESORES	?
◆ PROGRAMA	?
◆ REGLAMENTO	?
◆ SEA	?
◆ SER	?
◆ SERÁ	?
◆ SERÁN	?
◆ SIGUIENDO	?
◆ TENDRÁ	?
◆ TENDRÁN	?
◆ TESIS	?
◆ TIPO	?
◆ TUTOR	?
◆ ÁREA	?

Electronic lexicon of
nouns and verbs

Semantic Mining based on the Learning Graph

Example: websites analyze

Procesamiento de texto no estructurado

Introducir el Texto

La Universidad de Los Andes es una universidad pública y autónoma ubicada en los andes venezolanos con su sede principal y rectorado en la ciudad de Mérida; fundada por el clero como casa de estudios el 29 de marzo de 1785, elevada luego a seminario y finalmente reconocida como Universidad el 21 de septiembre de 1810 bajo decreto expedido por la Junta Gubernativa de la provincia de la Corona de España.

Es una de las principales universidades de Venezuela por la cantidad de estudiantes que alberga, por su nivel académico y por sus aportes en investigación que han contribuido al estudio y desarrollo de las ciencias. La universidad tiene como propósito fortalecer la formación integral iniciada en los ciclos de educación primaria y secundaria, además de formar equipos profesionales y técnicos necesarios para el desarrollo y progreso de Venezuela.

La universidad está conformada por 11 facultades repartidas en el Núcleo Mérida (ubicado en la ciudad de Mérida), 3 núcleos autónomos localizados en las ciudades de San Cristóbal, Trujillo y El Vígía, dos extensiones universitarias con estudios de pregrado, postgrado y actualización profesional en Tovar y en Valera, extensiones de actualización profesional en las ciudades de Barinas, Guanare, Barquisimeto, Maracaibo, Caracas, entre otras, y diversas instalaciones universitarias dentro del territorio nacional como estaciones experimentales, haciendas de producción agrícolas, reservas naturales para el desarrollo de la fauna y flora y laboratorios de investigación.

Separar Sentencia

Unstructured text



Reconocimiento de Entidades y Relaciones Candidatas

Entidades Candidatas

Universidad de Los Andes	universidad	andes	sede
ciudad	Mérida	universidades	Venezuela
cantidad	estudiantes	nivel	aportes
investigación	estudio	desarrollo	ciencias
propósito	formación	ciclos	educación
equipos	progreso	facultades	Núcleo
núcleos	ciudades	San Cristóbal	Trujillo
El Vígía	extensiones	estudios	pregrado
actualización	Tovar	Valera	Barinas
Guanare	Barquisimeto	Maracaibo	Caracas
instalaciones	territorio	estaciones	haciendas
producción	reservas	fauna	flora
laboratorios			

Relaciones Candidatas

es	Es	alberga	han
contribuido	tiene	fortalecer	formar
está			

Actualizar el Arbol

Universidad de Los Andes
Desarrollado por Tania Rodríguez y Inés Anzures

Entities and Relations Candidates

Semantic Mining based on the Learning Graph

Example: websites analyze

	Total
Candidate Entities	251
Candidate Relations	121

	Frecuencia	Peso	Porcentaje de Relevancia
Consejo Directivo	38	36,83	96,92
tesis	25	28,78	75,73
jurado	21	25,76	67,80
programa	21	25,76	67,80
tutor	21	25,76	67,80
Aspirante	18	23,29	61,28
doctorado	17	22,42	58,99
investigación	17	22,42	58,99
Plan de Formación	17	22,42	58,99
miembro	15	20,59	54,19
profesor	15	20,59	54,19
examen	14	19,64	51,68
créditos	12	17,64	46,42
caso	11	16,58	43,64
Area	10	15,49	40,76
estudiante	10	15,49	40,76
Estudios	10	15,49	40,76
grado	10	15,49	40,76
lapso	10	15,49	40,76
Programa de doctorado	10	15,49	40,76
artículo	9	14,35	37,77
comisión	9	14,35	37,77
conocimiento	9	14,35	37,77
doctoral	9	14,35	37,77
Facultad de Ingeniería	9	14,35	37,77
informe	9	14,35	37,77
postgrado	9	14,35	37,77
publicación	9	14,35	37,77
Año	8	13,17	34,65
candidatura	8	13,17	34,65
consejo	8	13,17	34,65
doctor	8	13,17	34,65
grupo	8	13,17	34,65
mes	8	13,17	34,65
reglamento	8	13,17	34,65
investigador	7	11,93	31,39
nível	7	11,93	31,39
tipo	7	11,93	31,39
Admisión	6	10,63	27,96
criterios	6	10,63	27,96
curso	6	10,63	27,96
país	6	10,63	27,96
Programa de doctorado en ciencias aplicadas	6	10,63	27,96
actividades	5	9,25	24,34

Relevant entities with weight criteria ≥ 10.06

build text summaries

Web Mining based on the Learning Graph

Example: Ontologies Construction

In the analyzed texts of the Doctorate website, we find the following sentences that have the verb 'to be'

The **doctoral student** is totally immersed in the dynamics of the research group to which his tutor belongs and follows the guidelines previously established by him in the training plan.

Any **qualified researcher** who is a member of a consolidated research group at the Universidad de Los Andes is, potentially, **a tutor** of the program.

If the **Training Plan** is not accepted by the Admission Committee, the applicant and his/her tutor can modify it and submit it once again to the Commission for consideration, within a period of one month.

Web Mining based on the Learning Graph

Example 2: Ontologies Construction

For each candidate sentence, the morphosyntactic analysis determines.

- In sentence 1 cannot be established a relationship among the entities.
- In sentence 2 is established the relationship that researcher is a tutor



- In sentence 3 cannot be established a relationship the verb is: "is accepted"

Ontology Mining

Ontology Mining

To explore techniques that can extract **additional knowledge from a set of ontologies**, to achieve a wider domain of knowledge.

- Extraction of knowledge patterns,
- Build or enrich ontologies.
- Establish relationships between ontologies
- ...

Ontology Mining

- **Extraction of Rules**
- **Integration of Ontologies**
 - **Linked Ontologies**
 - **Merge of Ontologies**
 - **Ontology Alignment**

Our works

Alignment of ontologies

Identify concepts of an ontology that are similar in the other ontologies

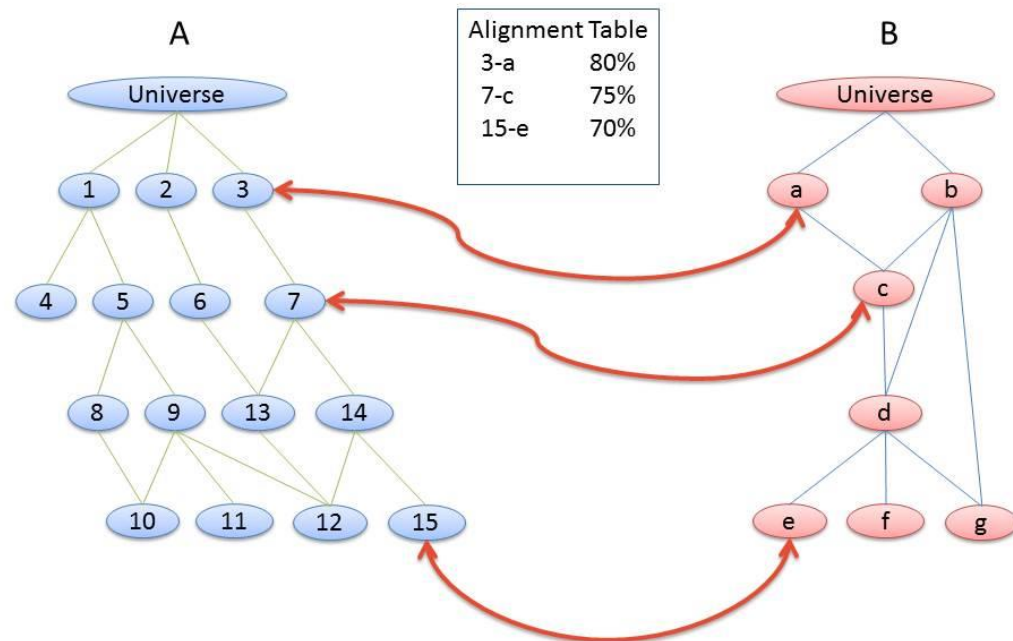
Semantic distance between each pair of concepts in different ontologies

Methods and tools for the alignment of ontologies

Alignment of ontologies

ontologies alignment techniques finds the correspondences between the concepts of the ontologies.

A & B Align



Alignment of ontologies

Ontology alignment techniques

- **Based on linguistic matching**
 - Similarity based on terms.
 - Semantic similarity:
 - Similarity between properties of the classes
 - Similarity between super-classes
- **Based on graph matching**

Distance of Levenshtein

Example of Method to Calculate Near Concepts

Assume that CA is a concept in the ontology A and PA its predecessor. CB is a concept CB in the ontology B (PB is its predecessor).

Four cases to calculate the similarity:

Case A: The CA concept coincides with CB in B, and the predecessors PA and PB

Case B: PA matches PB, but there is no match between CA and CB.

Case C: CA matches CB, but there is no match between PA and PB.

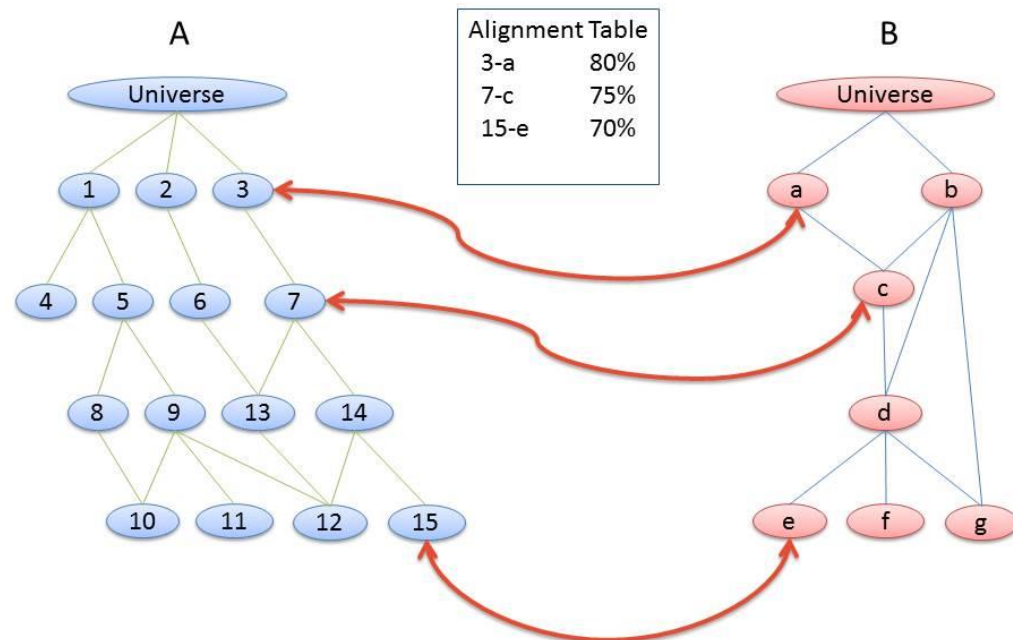
Case D: CA does not coincide with the CB and PA does not coincide with PB.

Ontology Alignment Recommendation System using ABC algorithm

The problem is to determine which alignment technique should be used at a given context.

We propose an **ABC** based technique, which **automatically select** the proper alignment **technique**

A & B Align



Emergent Alignment by using our ABC Approach

The gain $G(S_i)$ is calculated as follows, in the equation

$$G(S_i) = \frac{S_a(S_i)}{CA(S_i)} \times P_c$$

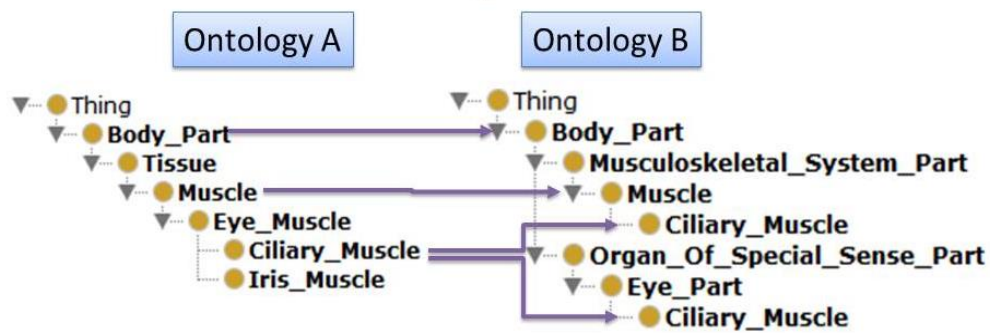
- **S_i :** Service that can be utilized to resolve a request (In our case, the alignment techniques). That is, **each alignment technique** is a source of nectar.
- **$G(S_i)$:** Profit, that is obtained by the use of the service S_j (one alignment technique), defined by the equation, which determines the **quality of nectar** (alignment technique).
- **$Sa(S_j)$:** Satisfaction of the Bee, when the service S_j is performed. It is also related with the quality of nectar; in our case is the **number of aligned nodes** of the ontologies
- **$CA(S_j)$:** Cost, it is represented in this work as **the execution time** of the service S_j to return a result (also affects the quality of nectar).
- **P_c :** Probability of preserving the food source. **Pseudo-random** value, with a **normal distribution** within the range of 0 and 1, which changes the value of $G(S_i)$

Used Alignment Techniques

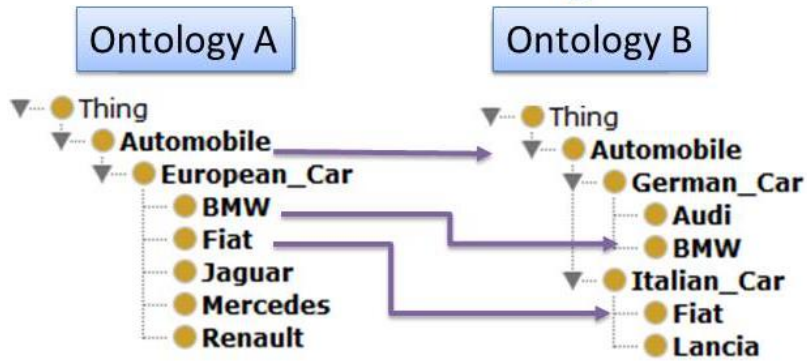
- a) **Class Structure:** based in finding similitude taking into consideration the **graph structure** from the classes.
- b) **Distance edited name:** semantic distance required to make **the classes names** equals.
- c) **Distance edited subclass name:** semantic distance required to make **the sub-classes names** equals, from the parent classes
- d) **Name and properties:** **similitude** between classes names and properties.
- e) **Same names:** semantic similitude if **names from classes** are equals.
- f) **Distance SMOA name** (A String Metric for Ontology Alignment): similitude between entities as **parts in common minus their differences**.
- g) **String Distance:** takes classes, properties and instances **as simples strings** to compare them.
- h) **Sub structures distance:** based in finding similitude taking into consideration the **graph structure** from the **sub-classes**.

Experiments

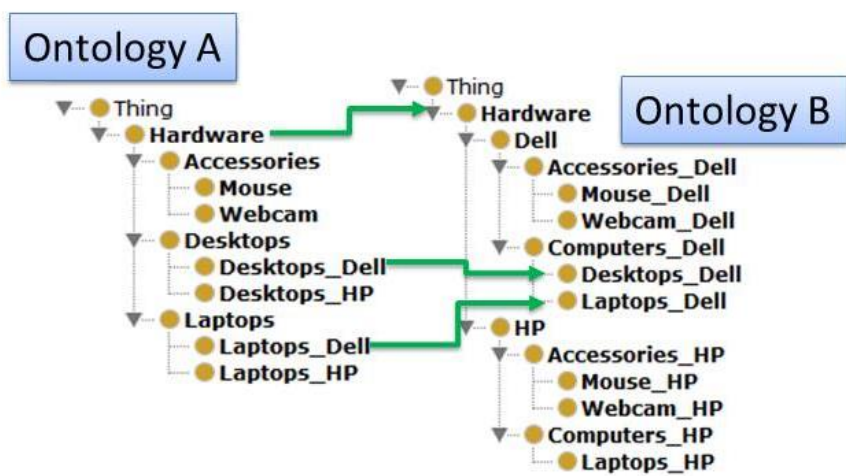
Anatomy Sub-set



Cars Ontologies



Computers Ontologies



Results

Set	Execution Time (sec)	# Aligned nodes	#Times that each Align Technique is chosen
Cars	1:00	0	a) 0
	0:50	3	b) 6
	0:51	3	c) 3
	0:49	3	d) 2
	0:41	3	e) 7
	0:40	3	f) 7
	0:59	3	g) 3
	1:12	3	h) 2
Anatomy Sub set	0:50	0	a) 0
	0:59	3	b) 0
	0:53	3	c) 0
	0:47	5	d) 12
	0:45	3	e) 0
	0:35	5	f) 8
	0:59	3	g) 0
	0:53	5	h) 10
Computers	1:02	0	a) 0
	1:06	4	b) 0
	0:56	7	c) 0
	0:49	9	d) 12
	0:49	4	e) 0
	0:49	9	f) 14

In this cases, all techniques from b) to h) align **the same numbers of nodes**, as the **times are closed**, it is chosen any of this techniques

In this two cases the number of aligned nodes are different, it is chosen any of the techniques with the bigger **quantity of aligned nodes**, choosing the one **with less time**.

Linking Ontologies

It is the process to find relationships between entities that belong to different ontologies.

Weak link of Ontologies: it is a correspondence between identical concepts.

Strong Link of Ontologies: It is carried out in a **semi-automatic** way, with the help of a global knowledge expert that is linking, which can define new concepts, as well as **links that relate concepts of different ontologies**, thus creating a **Meta-Ontology**.

Merge of ontologies

It is the process where several ontologies within the same domain come together to standardize knowledge, make knowledge grow or have total knowledge locally.

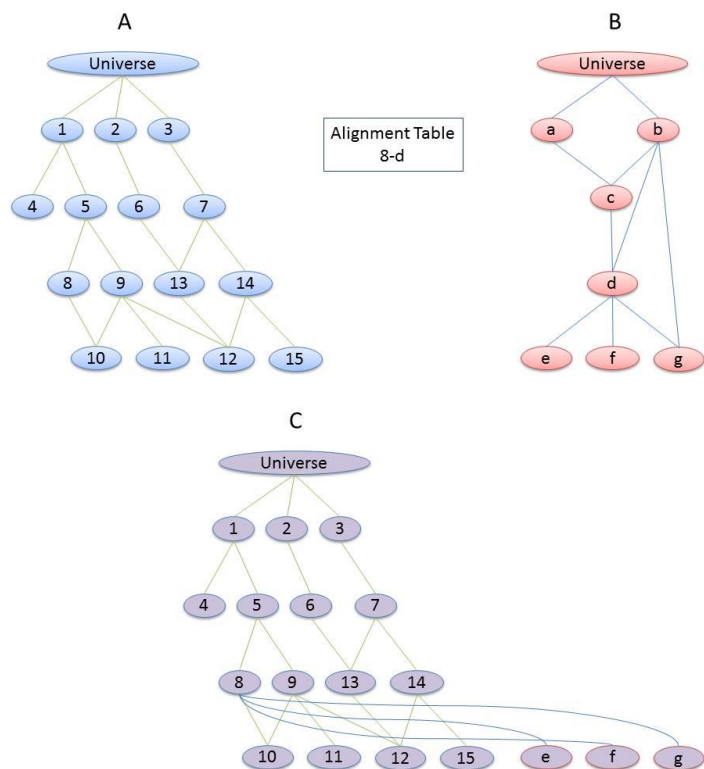
The ontologies handle the same knowledge, but with different representations, or have partial representations of that knowledge.

Weak mixture of Ontologies: take an ontology A, copy it as a result C, and enrich it with the other B, comparing all the concepts of ontology C (which are the same as A at this time) with those of ontology B, enriching the concepts of C with their similar concepts of B, **leaving out part of the knowledge of B.**

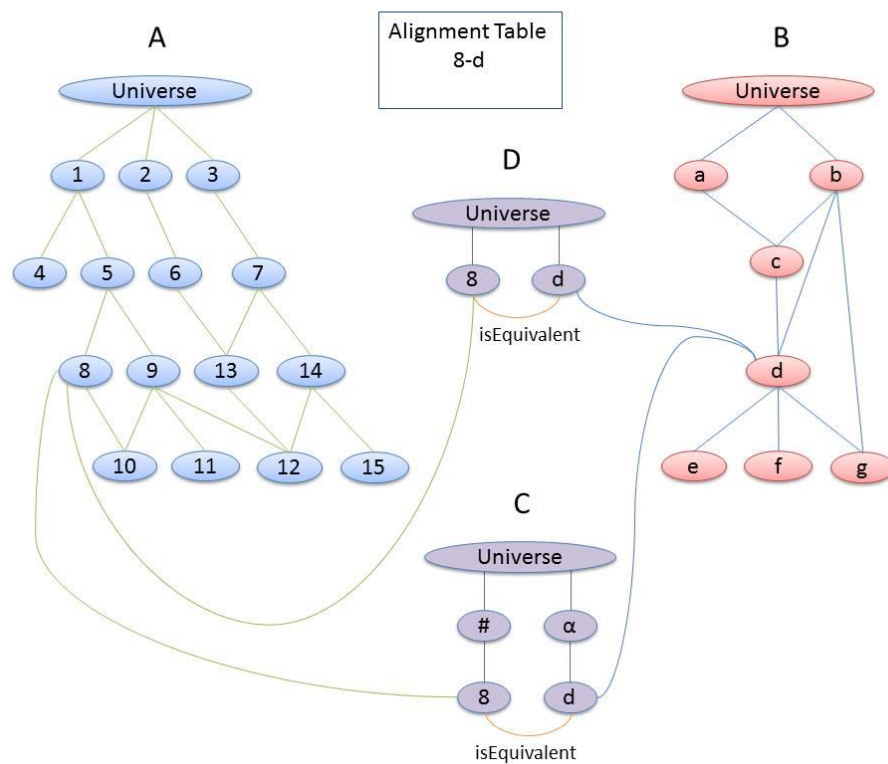
Strong mixture of Ontologies: It is a weak mixture, but incorporating the knowledge left out of B,

Ontology Mining (OM)

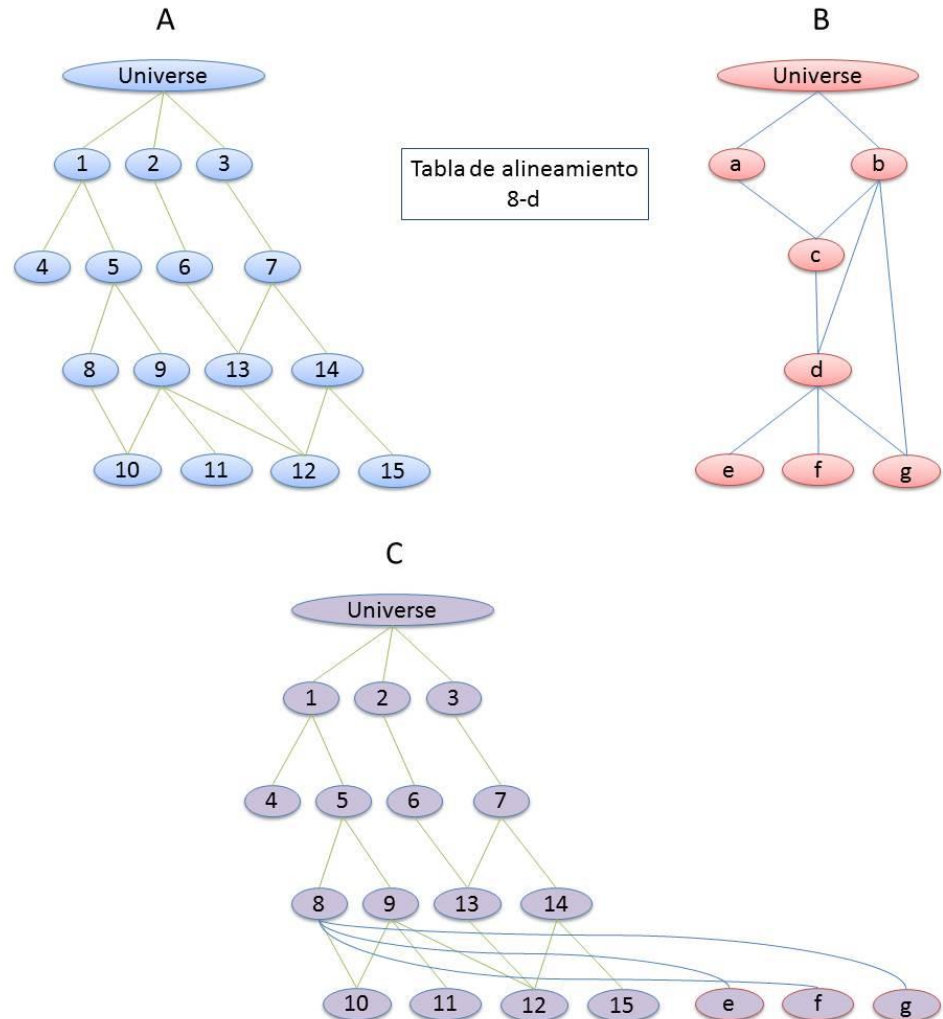
Merge of A & B



Linking of A & B



Ontology Merge Algorithms



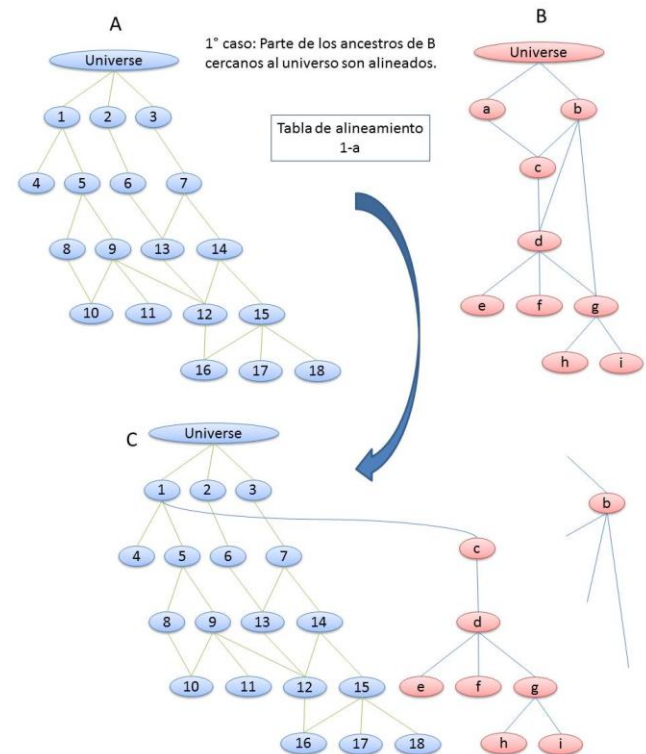
The problem of the **traditional mixture of ontologies** (which we have called **weak mixture**), is that it leaves knowledge without being incorporated in the resulting ontology.

Ontology Merge Algorithms

Our **Strong Blend** proposal is made in two parts,

- The weak mixture is made,
- The concepts and relationships left outside are incorporated.

First Part: Our system performs the weak mixture of two consistent ontologies A, B in an ontology C



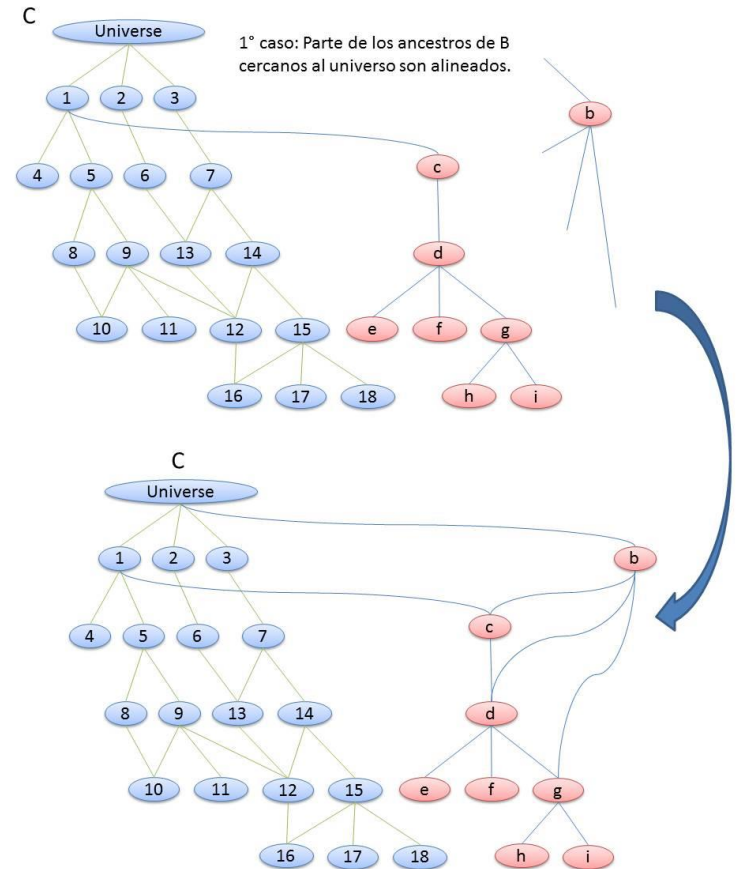
Ontology Merge Algorithms

Second Part: if the ontology to which the knowledge is being extracted to be added to the first has still **knowledge without being added**, the following cases are analyzed:

Case 1:

The concepts of B were **partially aligned** and **unaligned nodes are not copied** into the result ontology. Only would suffice with:

- add nodes not aligned to C
- copy the relationships that were not copied or aligned.



Ontology Merge Algorithms

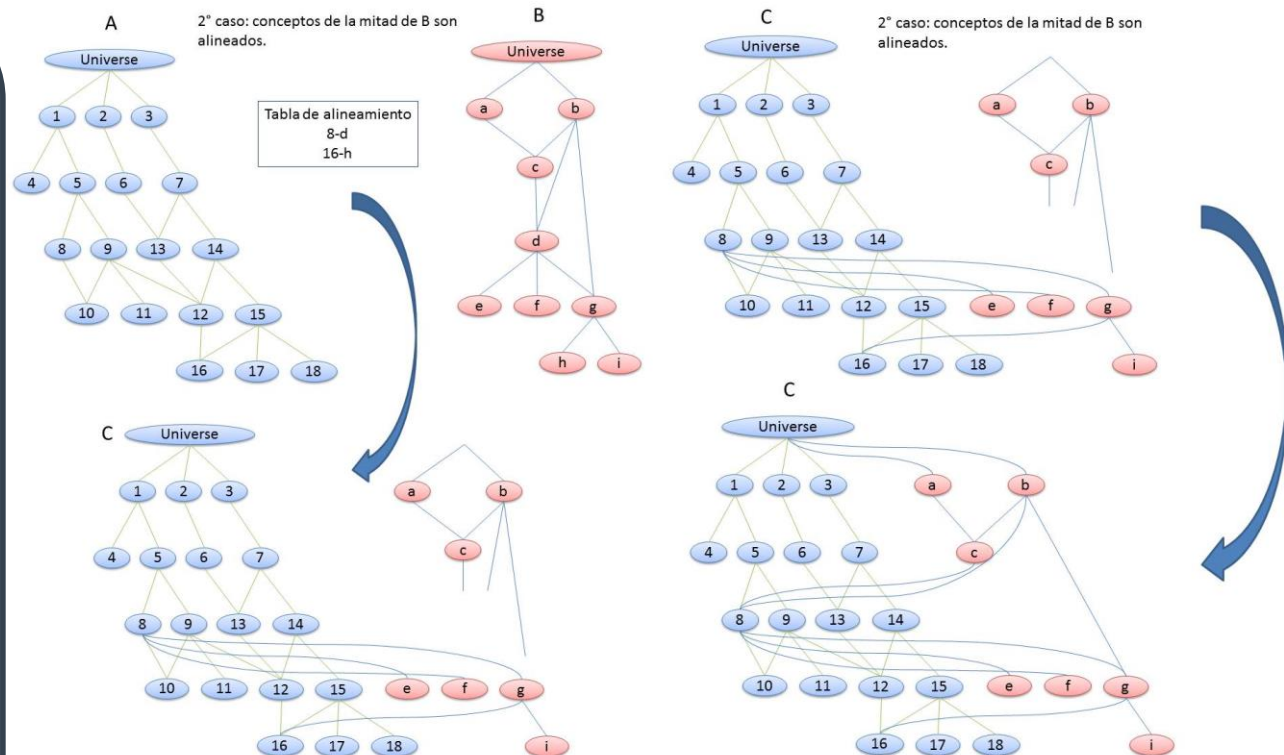
Case 2

- Certain intermediate nodes were aligned
- It leaves the ancestors without being copied

The strong mixture must:

- add these concepts to the universe of C
- Add the relationships where they participate

(see **b-g** ,**c-8** and **b-8**)



Ontology Merge Algorithms

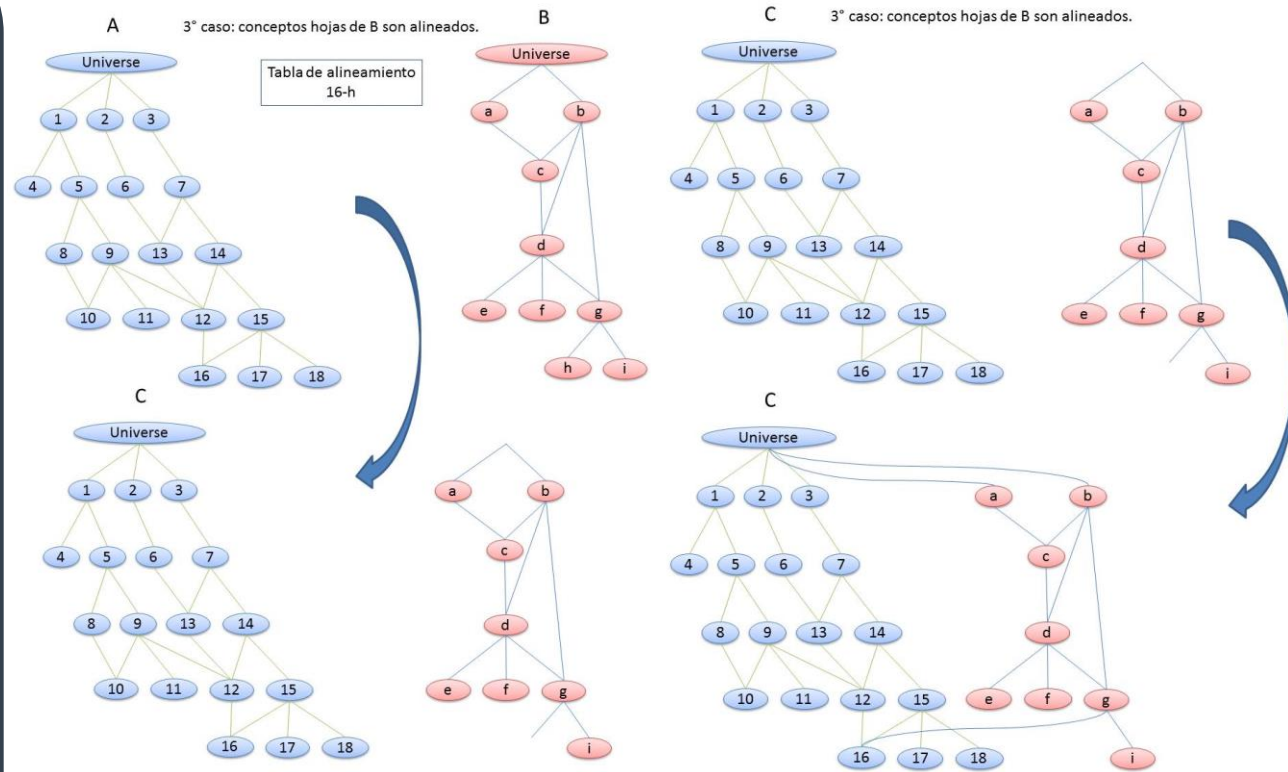
Case 3:

- Only the **leaf nodes** are **aligned**
- A large knowledge set of B. is left out of C.

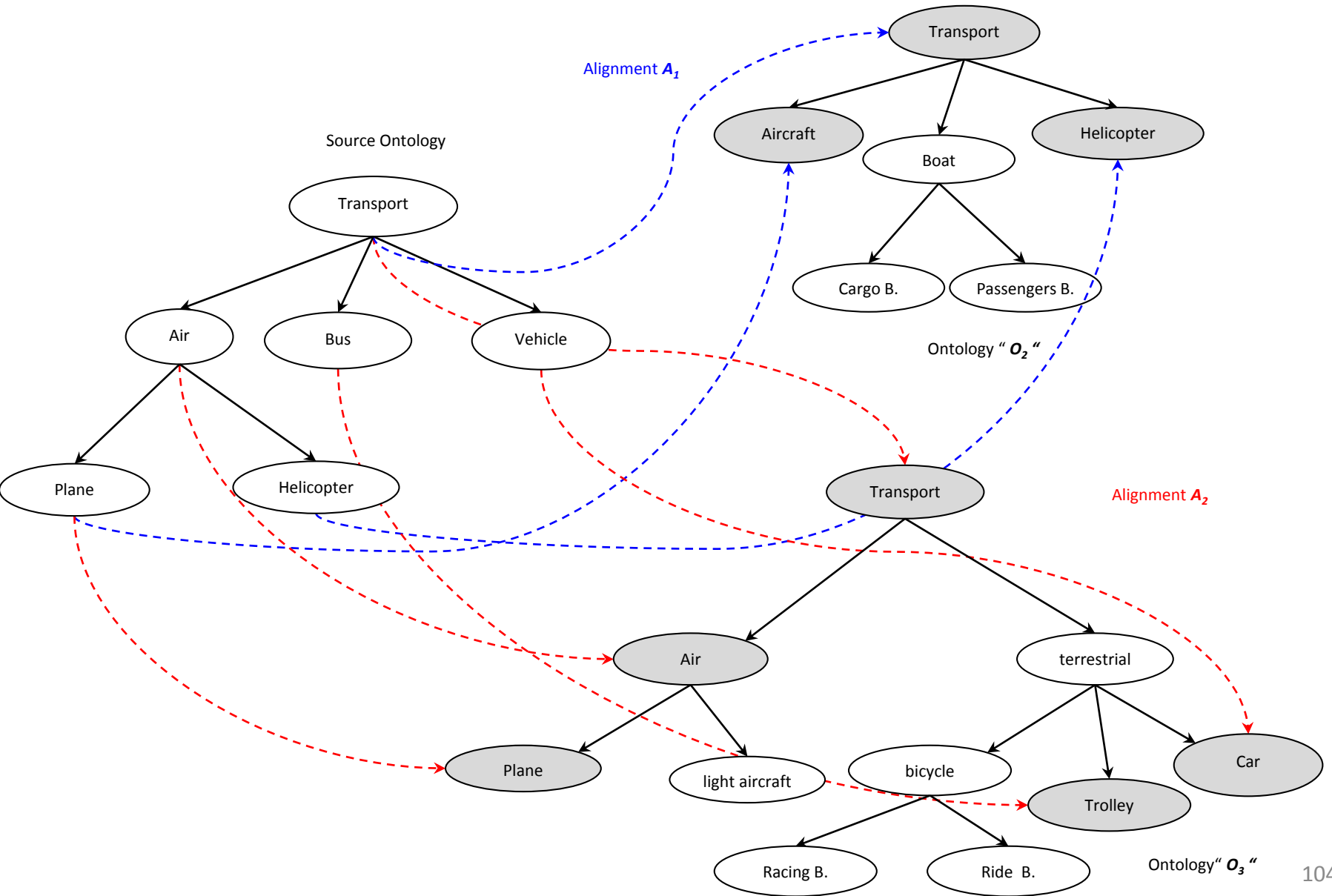
The strong mixture must

- **add** all concepts not copied to C
- **look** for the relationships that these concepts already had with others in B
- **copy** also those relationships with the concepts that were copied or with which they were aligned

(g-16)



Combination of multiple ontologies



Minería de Grafos

Jose Aguilar

CEMISID, Escuela de Sistemas

Facultad de Ingeniería

Universidad de Los Andes

Mérida, Venezuela

Grafos

Un grafo **G** es un par ordenado de un conjunto de vértices **V** y un conjunto de aristas **E**

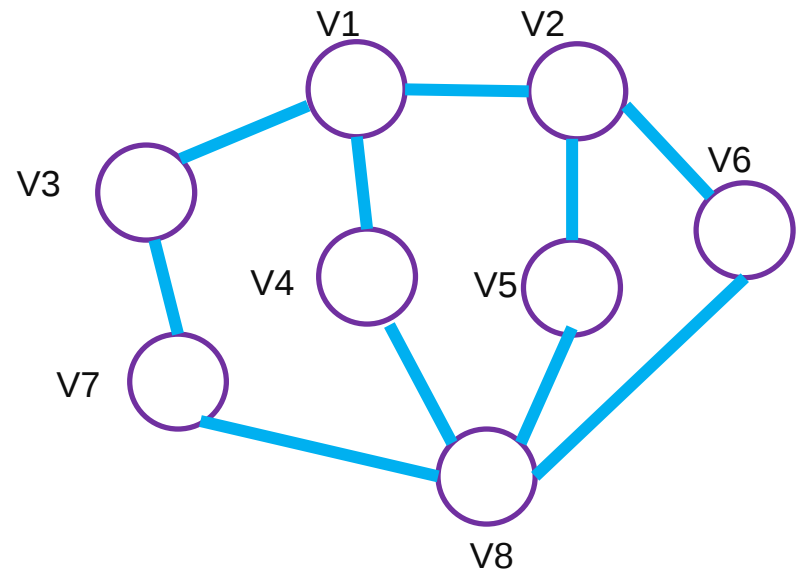
$$G = (V, E)$$

Par ordenado:

$(a, b) \neq (b, a)$ si $a \neq b$

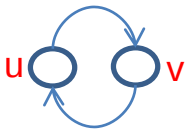
Par No ordenado:

$\{a, b\} = \{b, a\}$

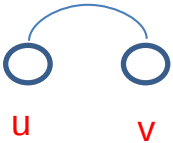


Grafos

Aristas:



Dirigido
(u, v)
(v, u)

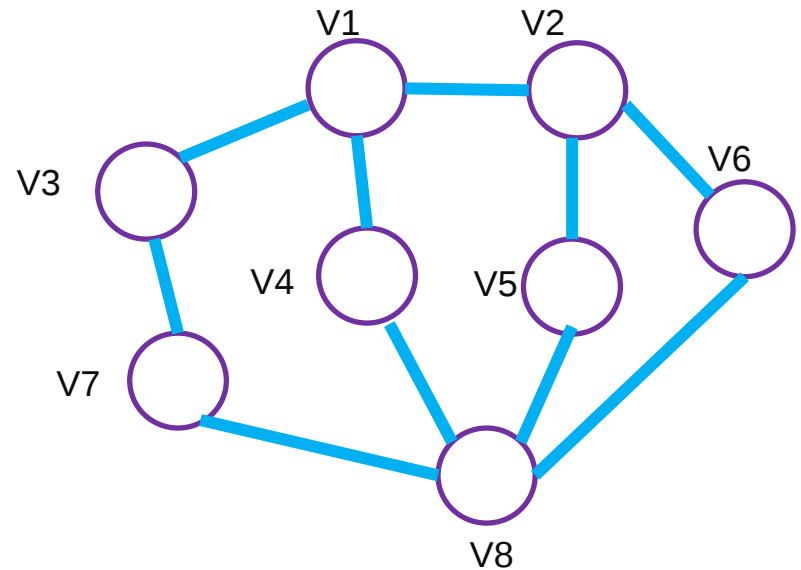
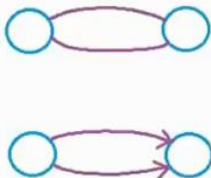


No Dirigido
 $\{u, v\}$

Bucle



Varias
Aristas



$V = \{V1, V2, V3, V4, V5, V6, V7, V8\}$

$E = \{\{v1, v2\}, \{v1, v3\}, \{v1, v4\}, \{v2, v5\}, \{v2, v6\}, \{v3, v7\}, \{v4, v8\}, \{v7, v8\}, \{v5, v8\}, \{v6, v8\}\}$

Grafos

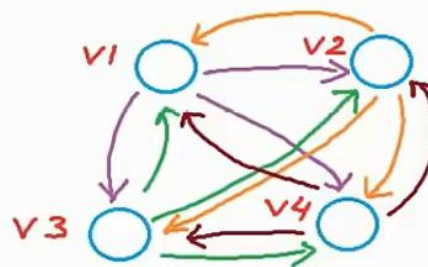
Número de
Aristas:

if $|V| = n$

then,

$0 \leq |E| \leq n(n-1)$, if directed

$0 \leq |E| \leq \frac{n(n-1)}{2}$, if undirected



if $|V| = 10$, $|E| \leq 90$

if $|V| = 100$, $|E| \leq 9900$

Modelando Datos con Grafos...

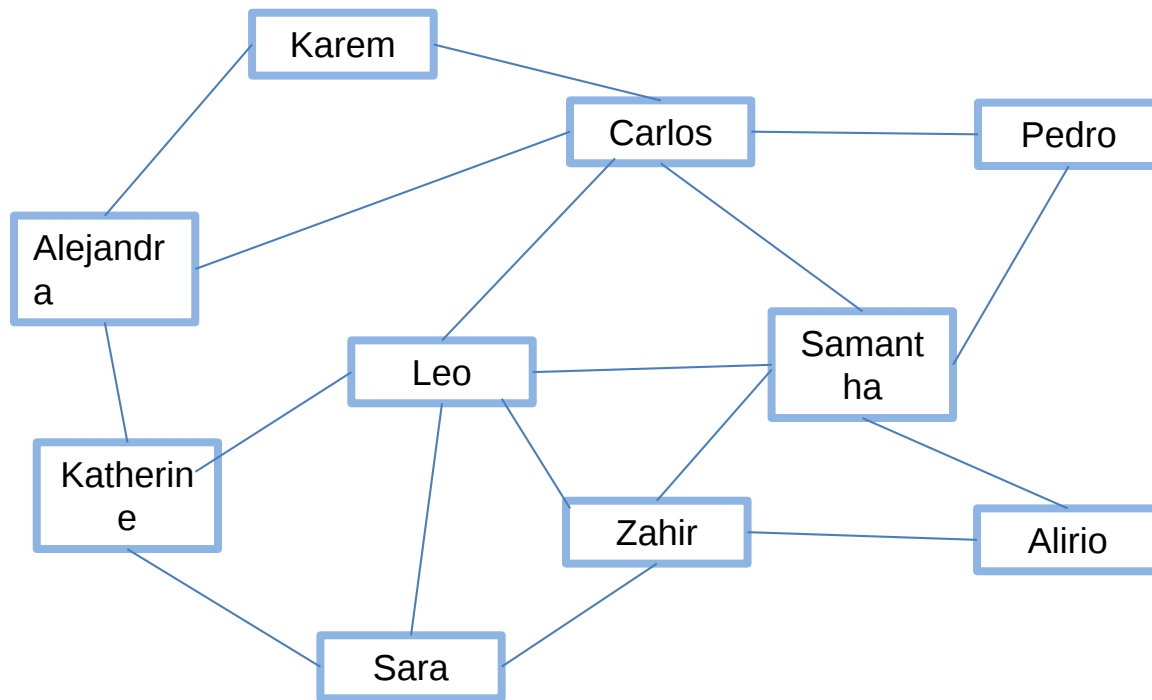
Los grafos son adecuados para la captura de las relaciones arbitrarias entre los diversos elementos.

<u>Instancia</u>		<u>Grafo</u>
Elemento	↔	Vertice
Atributos Elemento	↔	Etiquetas Vertices
Relaciones	↔	Arcos
Tipo de relaciones	↔	Etiquetas arcos

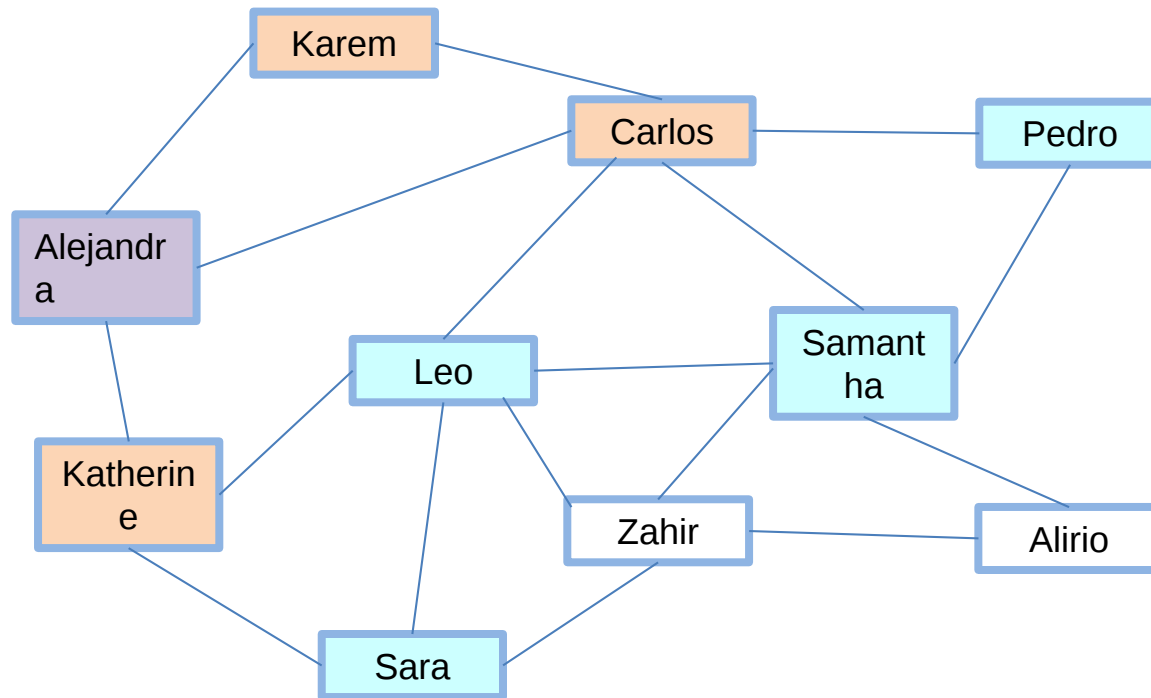
Proporcionan una enorme flexibilidad para el modelado de los datos, ya que permiten al modelador decidir cuáles son el tipo de relaciones a modelar

Grafos

Red Social
FACEBOOK



Grafos



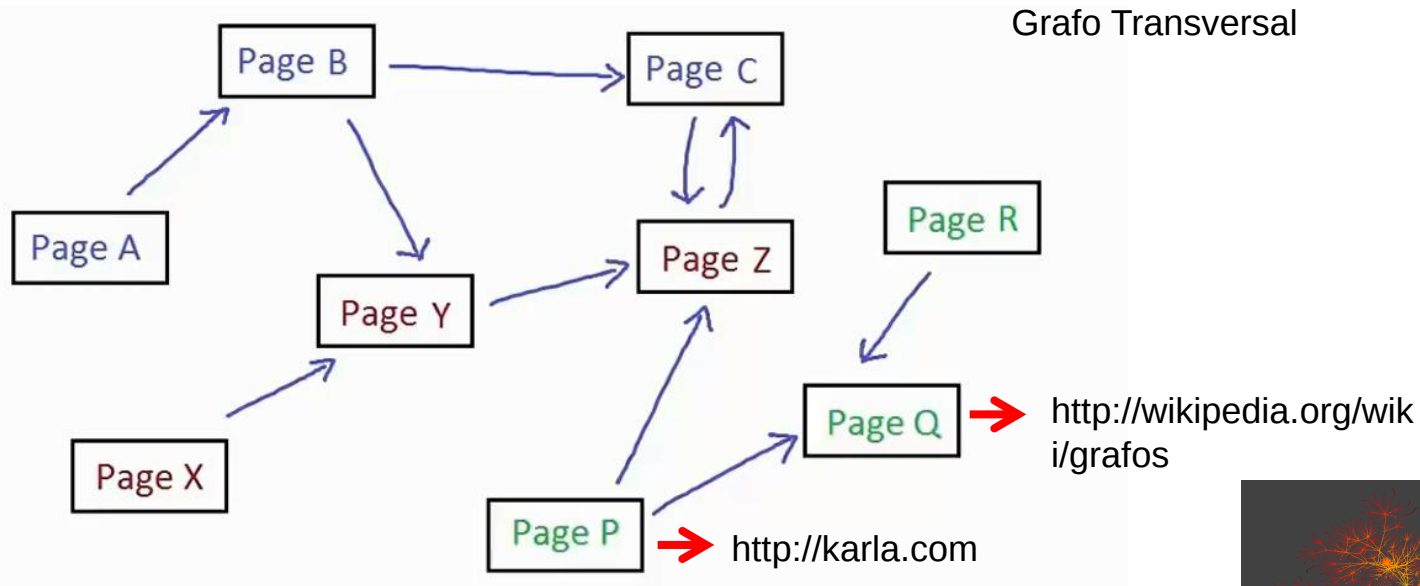
**Red Social
FACEBOOK**

Para Sugerir un amigo a ALEJANDRA hay que encontrar todos los nodos que tengan longitud del camino igual a 2.

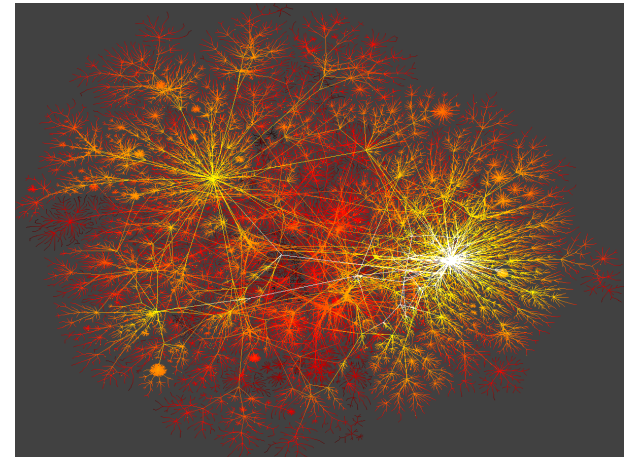
World Wide Web

Web – Crawling

Grafo Transversal

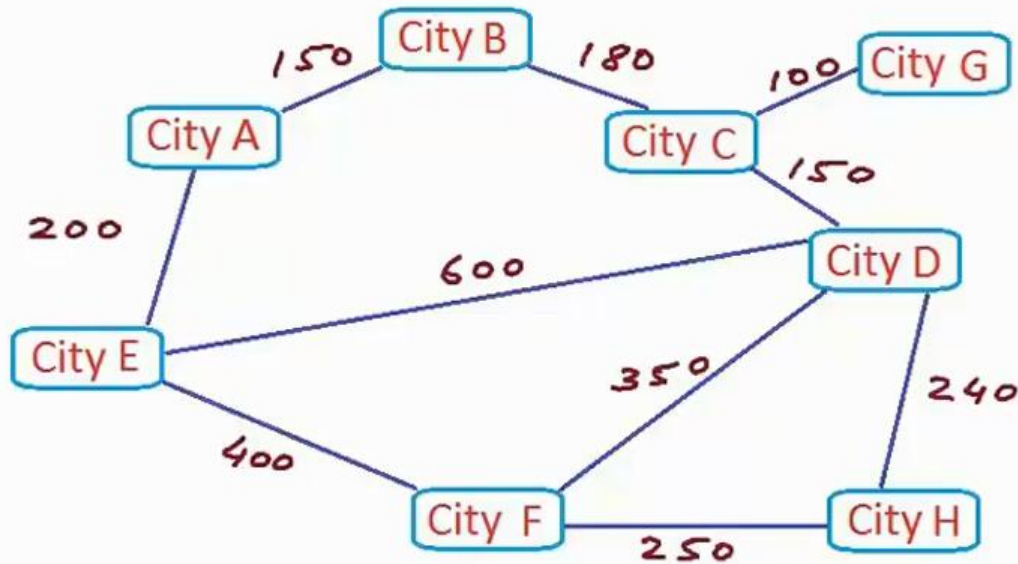


Internet



Grafos

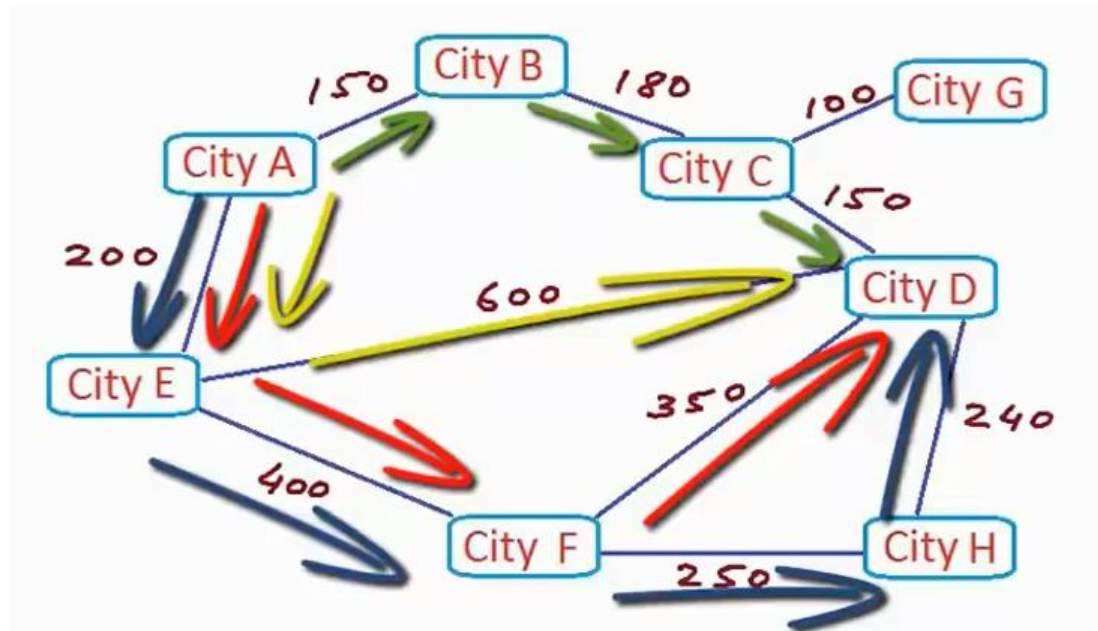
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Interurbanas

Grafos

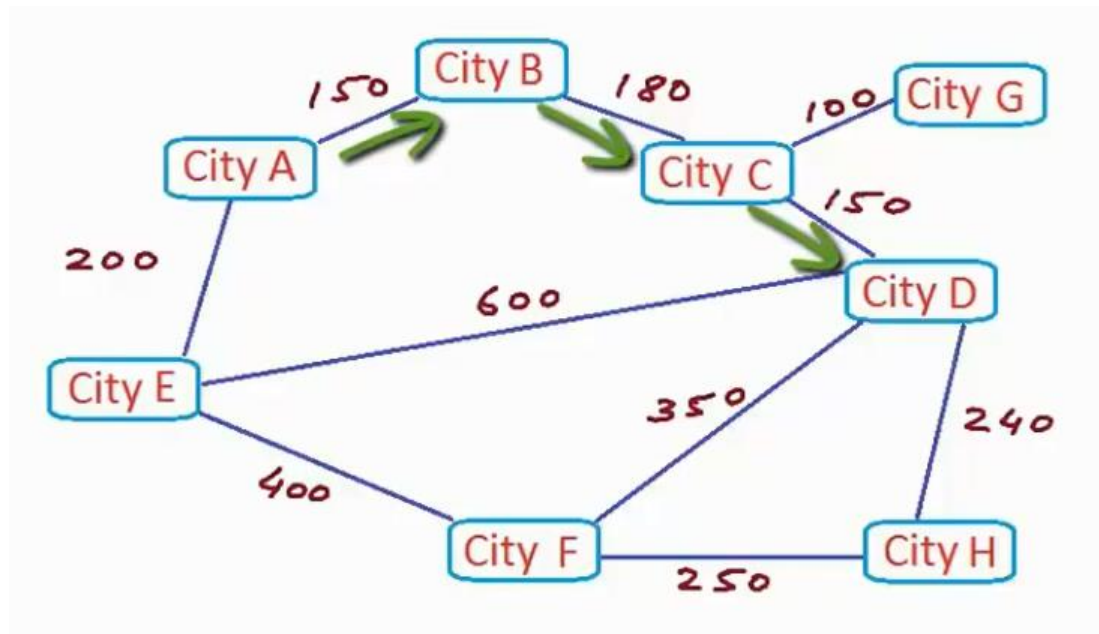
Grafos con Pesos VS Grafos sin Pesos



Red de Carreteras Interurbanas

Grafos

Grafos con Pesos VS Grafos sin Pesos



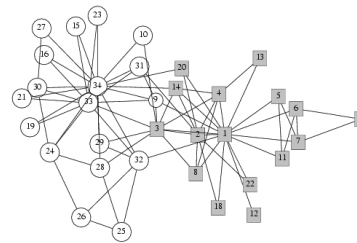
Red de Carreteras Interurbanas

Redes en el mundo real

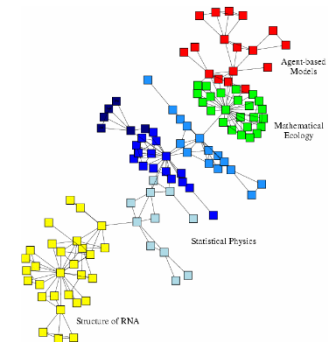
- **Redes de información:**
 - World Wide Web: hyperlinks
 - Redes de citación
 - Redes de Noticias y Blogs
- **Redes sociales**
 - Organizativas
 - Comunicativas
 - Colaborativas
 - Contactos sexuales
- **Redes tecnológicas:**
 - Energéticas
 - Transporte (aéreo, carreteras, fluviales,...)
 - Telefónicas
 - Internet
 - Sistemas Autónomos



Redes de amistad



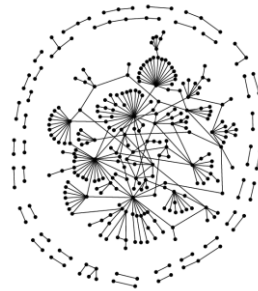
Karate club network



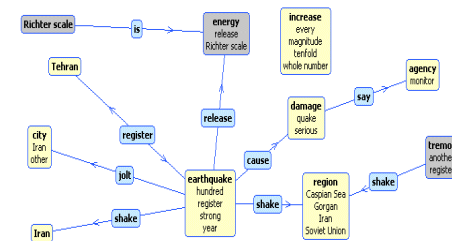
Redes de colaboración

Redes en el mundo real

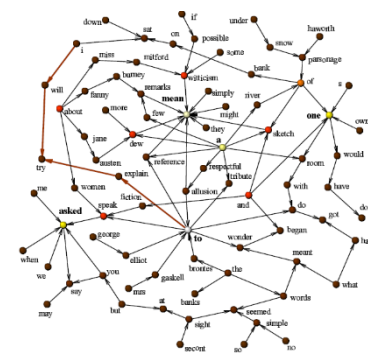
- Redes biológicas
 - Metabólicas
 - Cadenas alimenticias
 - Neuronales
 - Regulación Genética
- Redes de lenguaje
 - Semánticas
 - Lingüísticas
- Redes de software
- ...



Interacciones entre las proteínas de la levadura

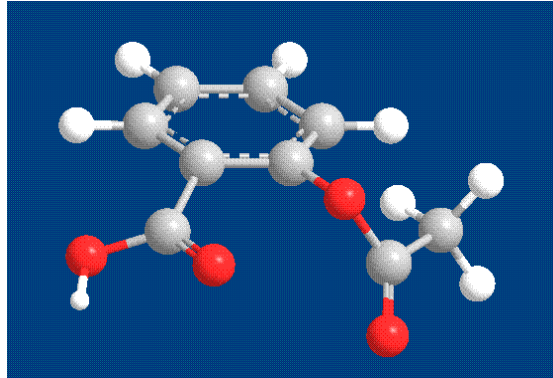


Red semántica

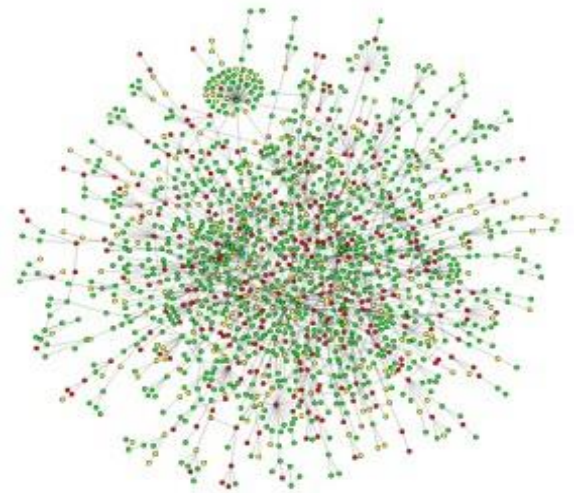


Red Lingüística

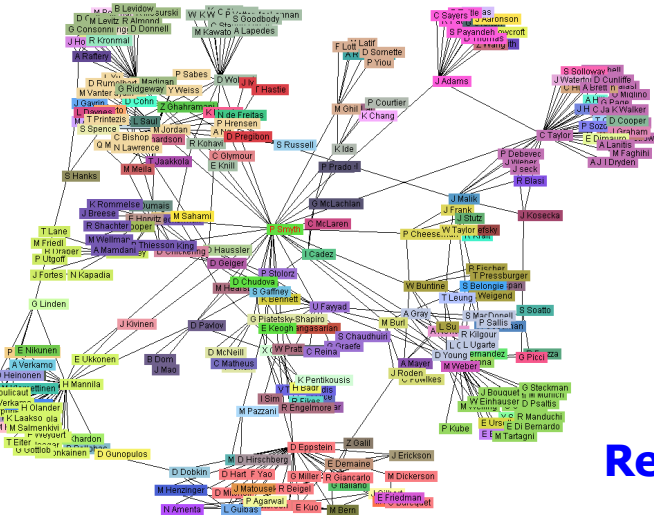
Grafos



Aspirina



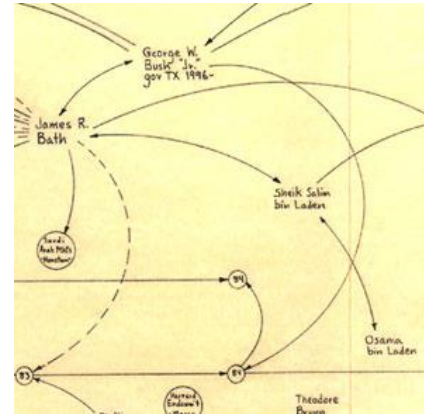
Red de interacción de proteína levadura



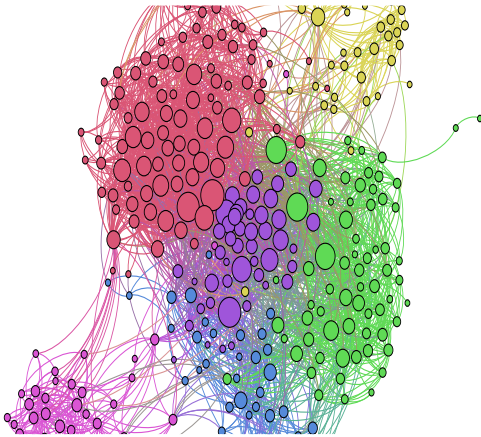
Red Co-autores de libros

Grafos

Mark Lombardi: rastreo y Mapeo fiascos financieros globales en los años 1980 a partir de fuentes públicas, como los artículos de noticias.

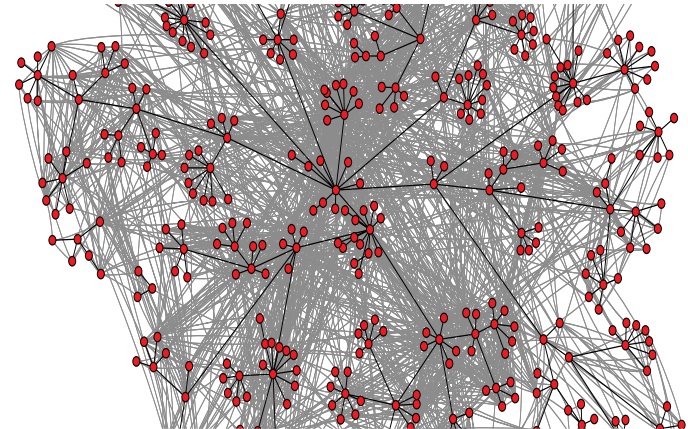


Relación de empleados de una organización



Facebook de alguien

Los colores separan componentes fuertemente conectados de la red.



Los arcos negros denotan estructura organizacional y los grises son interacciones por correo electrónico.

Red Social

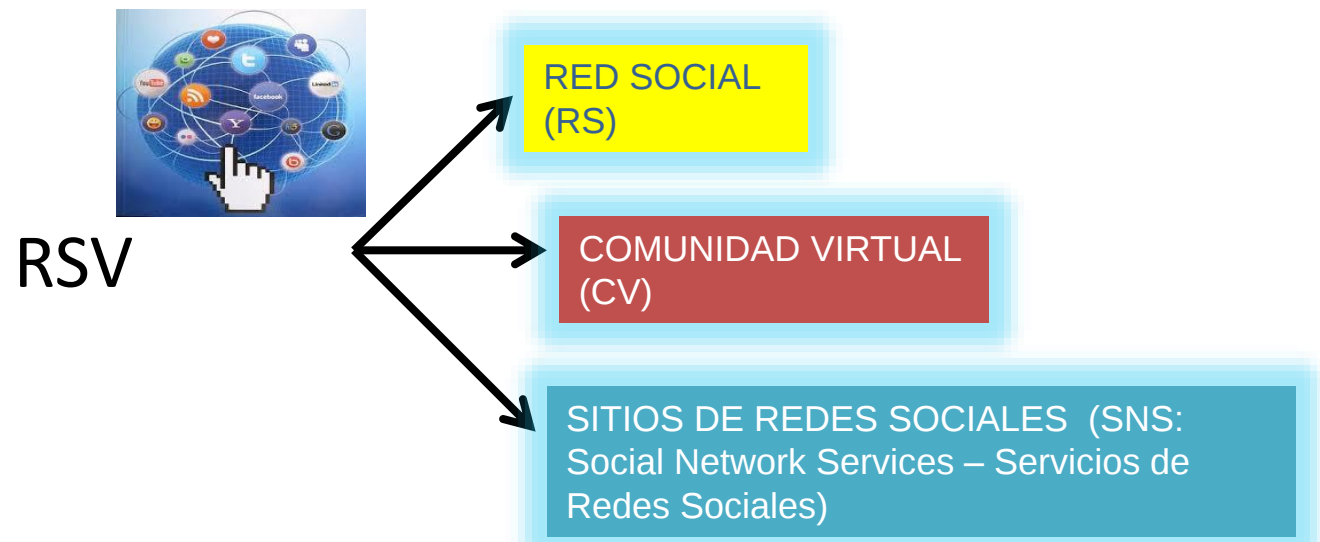


- El **análisis de redes sociales** estudia esta estructura social aplicando la teoría de grafos.
- Se analiza:
 - Si existen estructuras de comunidades ocultas
 - La difusión o las opiniones.
 - La influencia del todo en las partes y viceversa.
 - La difusión de nuevas ideas y prácticas (teoría de difusión de innovaciones).
 - El efecto producido por la acción selectiva de los individuos en la red
 - Grafos de colaboración para ilustrar buenas (amistad, alianza, citas) y malas (odio, ira) relaciones entre los seres humanos.

RED SOCIAL VIRTUAL WEB 2.0 Y 3.0



- **Facebook** es la red social más utilizada del mundo
- **Twitter:** red social de microblogging.
- **LinkedIn** red de usuarios profesionales, y
- **Youtube** red de alojamiento de vídeos.
- **Google+**, apuesta de Google por las redes sociales.
- **Instagram**, red de intercambio de imagenes



Herramientas

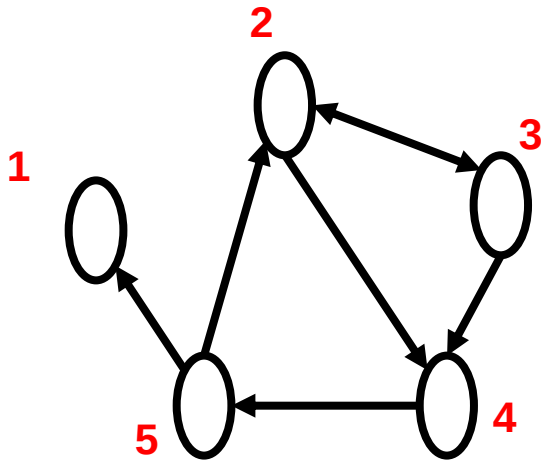
- **Gephi** (visualization and basic network metrics)
- **NetLogo** (modeling network dynamics)
- **Pajek**: amplia funcionalidad basada en menús, incluyendo muchas, muchas métricas de red y manipulaciones
 - pero ... no extensible
- **Guess**: extensibles, herramientas de secuencias de comandos de análisis exploratorio de datos, pero la selección más limitada de métodos incorporados en comparación con Pajek
- **NetLogo**: plataforma general agente basado en la simulación con el apoyo de modelado excelente red
 - muchos de los demos en este curso fueron contruidos con NetLogo
- **IGRAPH**: utilizado en la versión de nivel de doctorado. bibliotecas se puede acceder a través de R o Python. Rutinas escalan a millones de nodos. (for programming assignments)

Elementos de un grafo

- Dirigido
 - ▣ $A \rightarrow B$
 - A le gusta B, A le dio un regalo a B, A es hijo de B
- No dirigido
 - ▣ $A \leftrightarrow B$ o $A - B$
 - A y B se gustan, son semejantes
 - ▣ Peso (frecuencia de comunicación)
 - ▣ ranking (mejor amigo, segundo mejor amigo...)
 - ▣ tipo (amigo, pariente, co-trabajador)

Representación de los datos

Matriz de adyacencia



$A =$

0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	0	0	0	1
1	1	0	0	0

Lista de adyacencia

Todos los vecinos de cada nodo

- 1:
- 2: 3 4
- 3: 2 4
- 4: 5
- 5: 1 2

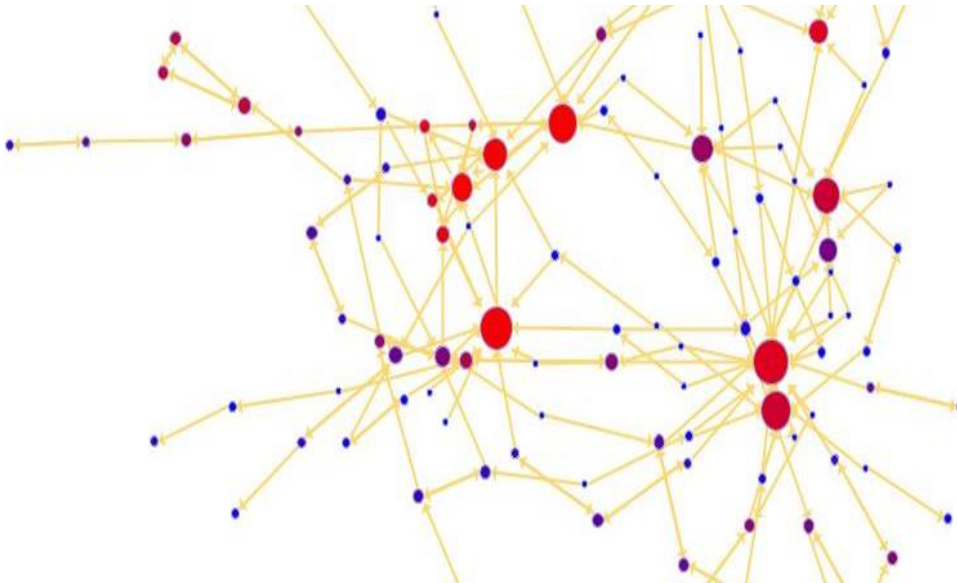
Lista de arcos

- 2, 3
- 2, 4
- 3, 2
- 3, 4
- 4, 5
- 5, 2
- 5, 1

Más fácil para redes

- Grandes
- Dispersas

Métricas



¿Cuál es el nodo con más arcos?

Métricas de redes

Cada métrica de red da respuesta a las siguientes preguntas:

➤ pregunta: ¿Quién es más central?

1) METRICA DE RED: centralidad

a) Centralidad de grado (degree centrality).

1) Indegree o grado de entrada

2) Outdegree o grado de salida

b) Centralidad de cercanía (closeness centrality).

c) Centralidad de intermediación (Betweenness centrality).

➤ pregunta: ¿Todo está conectado?

2) METRICA DE RED: los componentes conectados

- Componentes fuertemente conectados:

-Componentes Débilmente conectados:

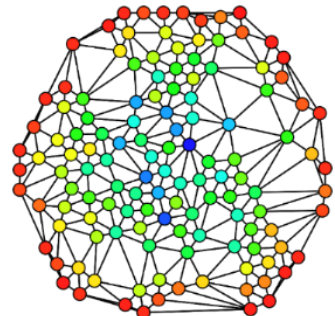
3) METRICA DE RED: tamaño de componente gigante(giant component)

➤ pregunta: ¿A qué distancia están las cosas?

4) METRICA DE RED: rutas más cortas

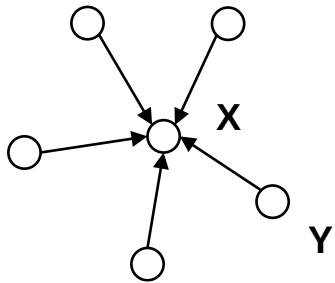
➤ pregunta: ¿Cómo densa son?

5) METRICA DE RED: densidad grafo

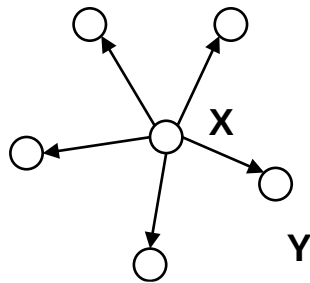


Métricas: Centralidad

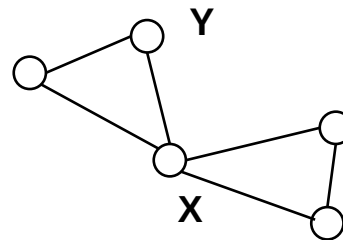
Medidas posibles de un vértice en un grafo, que determina su importancia relativa dentro de éste



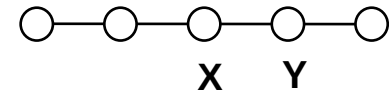
indegree



outdegree



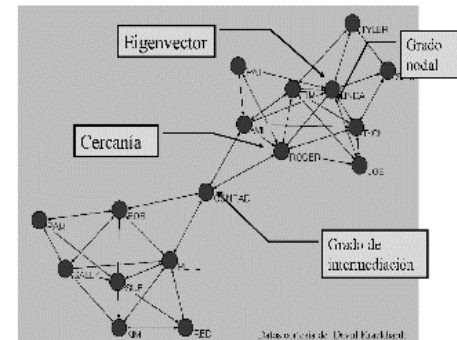
Betweenness
(intermediación)



Closeness
(cercanía)

Centralidad de vector propio (eigenvector centrality)

Cuatro Aspectos de la Centralidad



Métricas: Propiedades de los nodos de la Red

▣ Conexiones

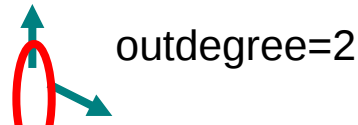
■ **indegree**

cuantos arcos están dirigidos al nodo



■ **outdegree**

arcos que salen del nodo



■ **degree (in or out)**

todos los arcos del nodo, entrada y salida



$$\sum_{i=1}^n A_{ij}$$

$$\sum_{j=1}^n A_{ij}$$

□ **Degree sequence:** Lista ordenada de los grados de cada nodo

■ In-degree sequence:

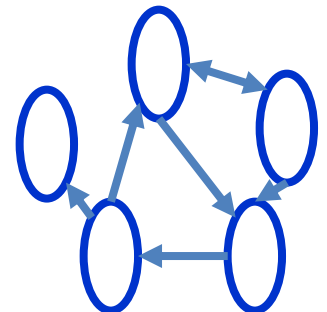
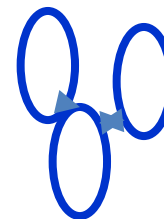
■ [2, 2, 2, 1, 1, 1, 1, 0]

■ Out-degree sequence:

■ [2, 2, 2, 2, 1, 1, 1, 0]

■ (undirected) degree sequence:

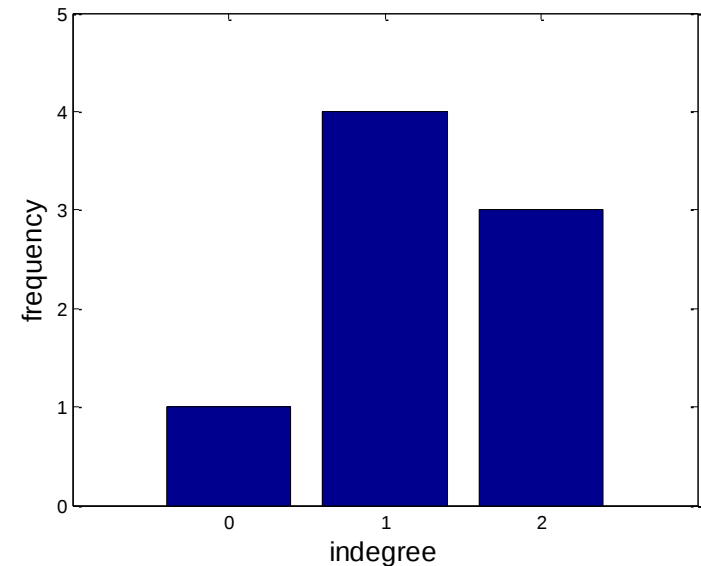
■ [3, 3, 3, 2, 2, 1, 1, 1]



Métricas: Propiedades de los nodos de la Red

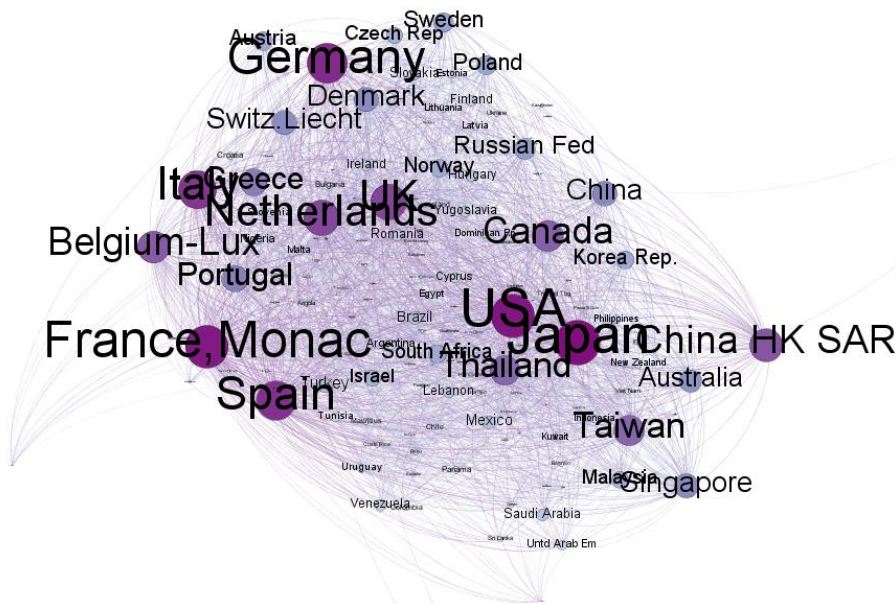
- **Degree distribution:** La frecuencia con la que ocurre cada grado

- In-degree distribution:
 - [(2,3) (1,4) (0,1)]
- Out-degree distribution:
 - [(2,4) (1,3) (0,1)]
- (undirected) distribution:
 - [(3,3) (2,2) (1,3)]

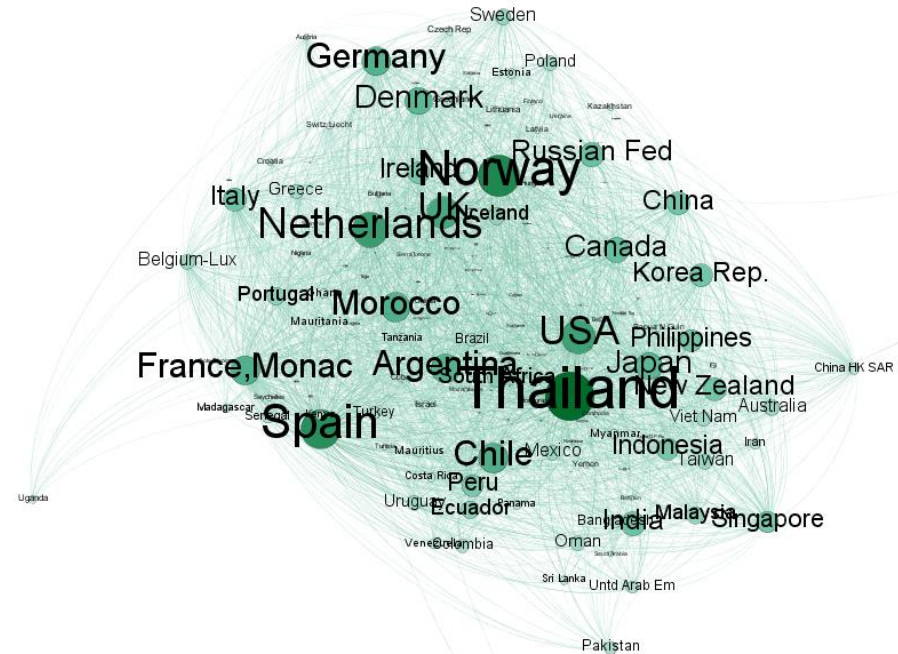


Métricas

InDegree



OutDegree



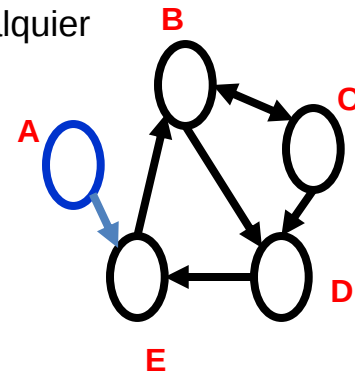
Métricas: Componentes conectados

Componentes fuertemente conectados:

- Cada nodo dentro del componente se puede llegar desde cualquier otro nodo en el componente siguiendo los enlaces dirigidos

- Componentes fuertemente conectados

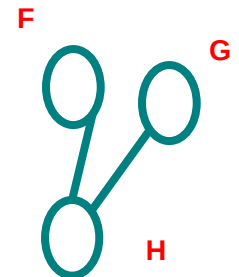
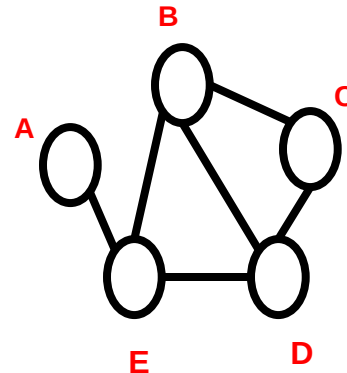
- B C D E
- G H



Componentes Débilmente conectados: cada nodo se puede llegar desde cualquier otro nodo siguiendo enlaces en cualquier dirección

- Componentes débilmente conectados

- A B C D E
- G H F



En las **redes no dirigidos** se habla simplemente de "**componentes conectados**"

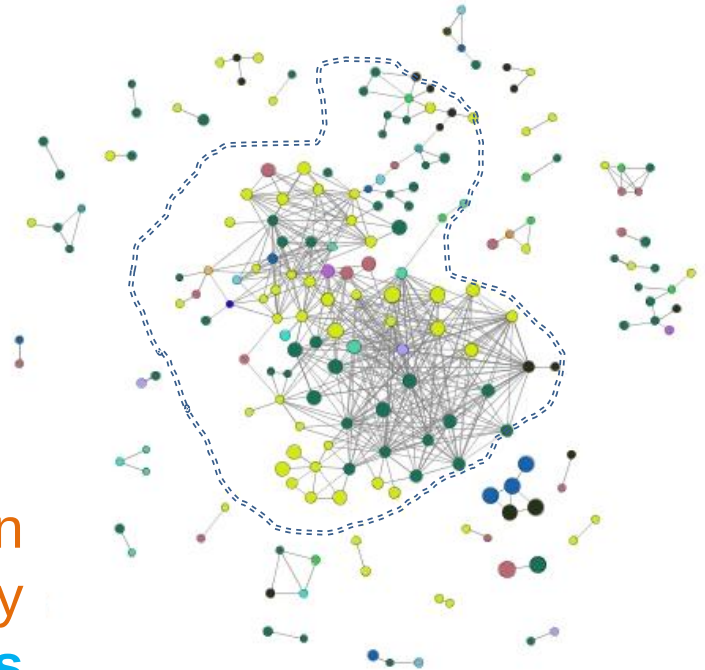
Métricas: Componentes conectados

- Si el componente más grande ocupa una región significativa de la red o grafo, es llamado

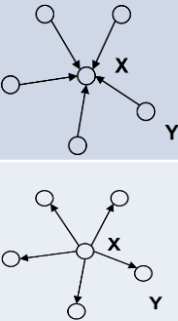
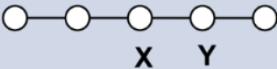
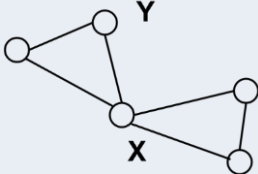
giant component

El componente gigante, consiste en un **grupo de nodos enlazados entre si**, y que agrupan a la **mayoría de los nodos** de la red.

El componente gigante aparece en casi **todas las redes sociales**



Métricas principales de la centralidad

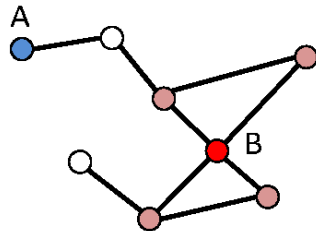
Métricas de centralidad		
a) Centralidad de grado (<i>degree centrality</i>): qué tantas conexiones directas tiene una unidad con otras unidades. Una unidad con alta centralidad de grado sirve como “conector” o “hub” de la red.	Indegree o grado de entrada outdegree o grado de salida	
b) Centralidad de cercanía (<i>closeness centrality</i>): que tan “cerca” se encuentra una unidad de la red de las otras, considerando tanto conexiones directas como indirectas. una unidad con alta centralidad de cercanía puede interactuar fácilmente con otras unidades, tiene la visibilidad del comportamiento de la red en su conjunto, y puede influir en ella.	closeness o cercanía	
c) Centralidad de intermediación (<i>Betweenness centrality</i>): índice de en qué tantas rutas más cortas entre 2 unidades cualesquiera de la red se encuentra una unidad dada. Estas unidades tienen el control del flujo de información dentro de la red.	betweenness o intermediación	

Métricas: Centralidad

MEDIDAS LOCALES DE CENTRALIDAD

Centralidad de grado

No dirigida

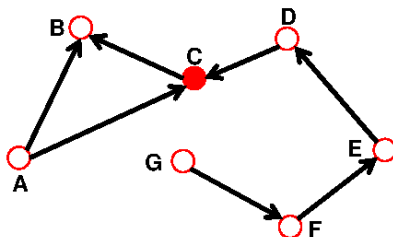


Centralidad de grado de un actor (C_D): número de enlaces que lo conectan con otros

$$C_D(\text{A}) = k_{\text{A}} = 1 \quad C_D(\text{B}) = k_{\text{B}} = 4$$

$C_D(i)$ se define en $\{0, g-1\}$, siendo g el número de nodos de la componente conexa

Dirigida



En redes dirigidas, se define el **Prestigio** de entrada (*in-degree*), denominado **Soporte**, y el **Prestigio** de salida (*out-degree*), denominado **Influencia**:

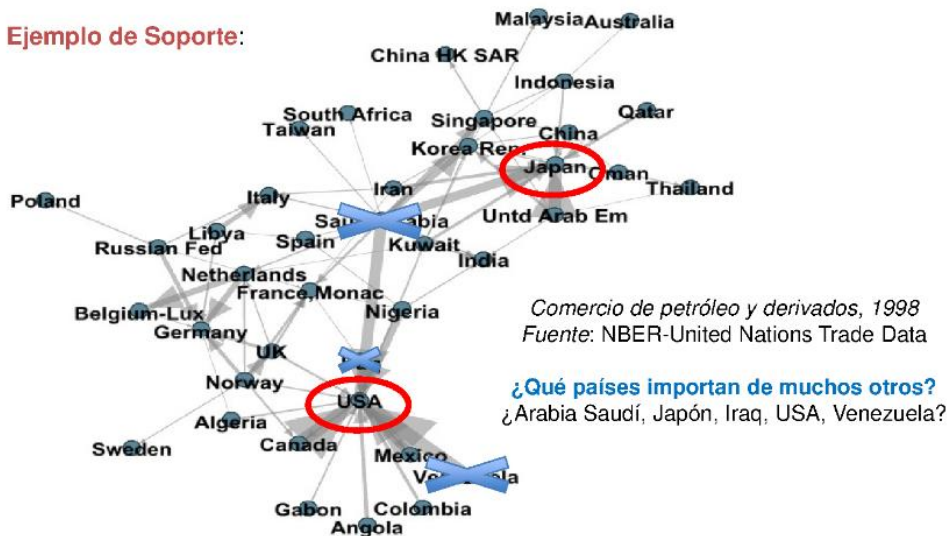
$$P_D^{in}(\mathbf{C}) = k_C^{in} = 2 \quad P_D^{out}(\mathbf{C}) = k_C^{out} = 1$$

Ambos se definen en $\{0, g-1\}$

MEDIDAS LOCALES DE CENTRALIDAD

Soporte

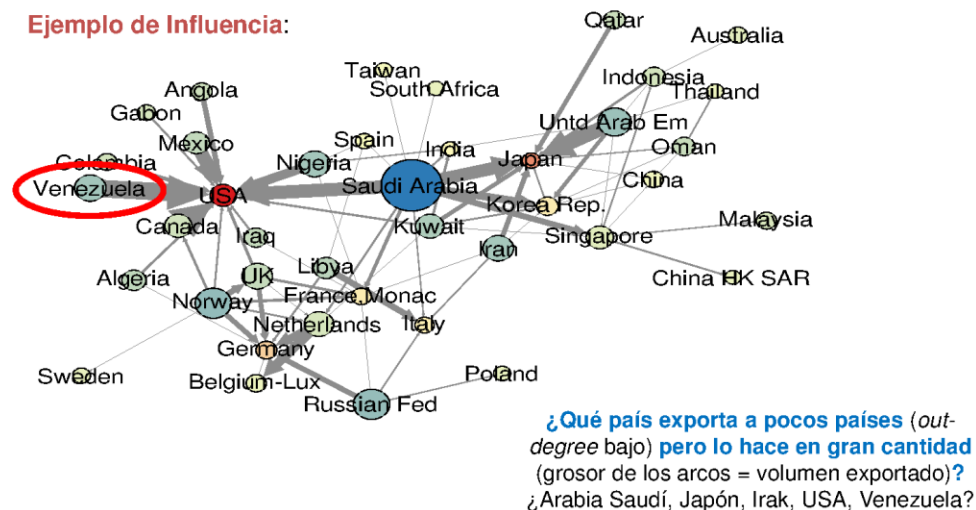
Ejemplo de Soporte:



MEDIDAS LOCALES DE CENTRALIDAD

Influencia

Ejemplo de Influencia:



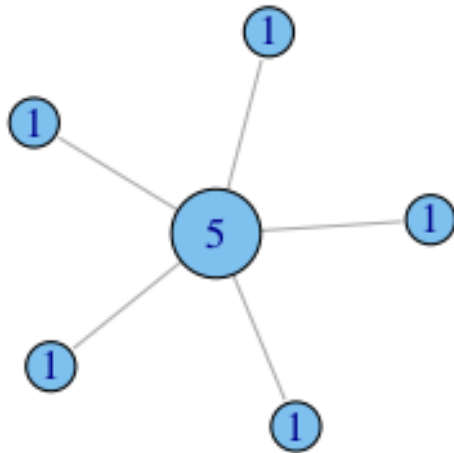
Métricas: Centralidad

$$C'_D(p_k) = \frac{\sum_{i=1}^n a(p_i, p_k)}{n-1}$$

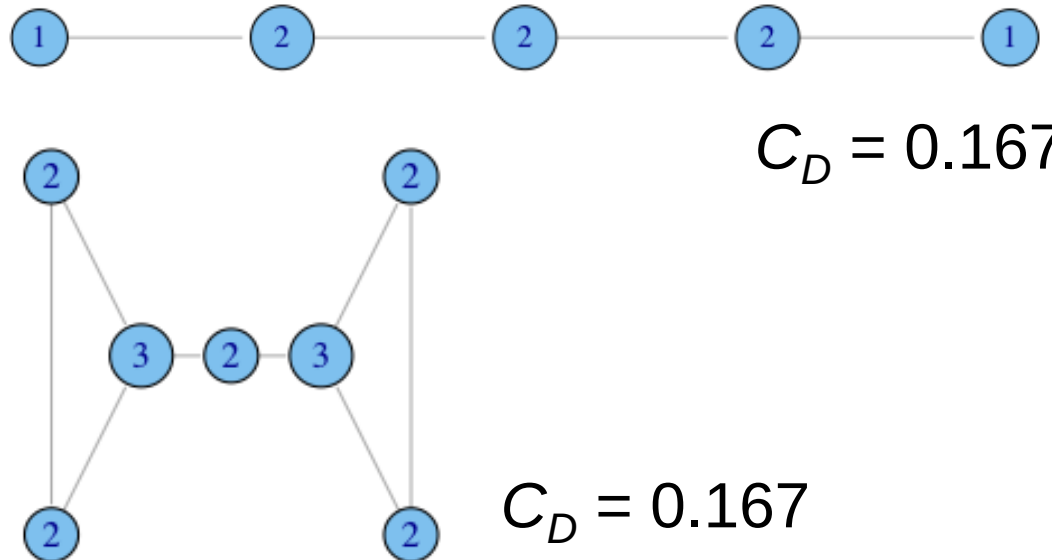
$$C_D = \frac{\sum_{i=1}^n [C_D(n^*) \cdot C_D(i)]}{[(N-1)(N-2)]}$$

Máximo valor de conexiones posibles en la red

Formula de centralidad general de Freeman's



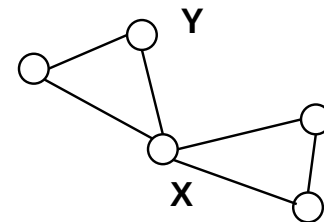
$$C_D = 1.0$$



$$C_D = 0.167$$

$$C_D = 0.167$$

Métricas: betweenness



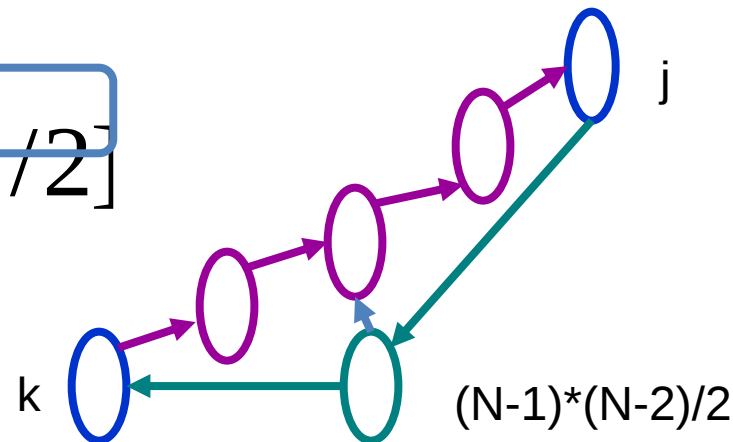
La centralidad de intermediación ve al nodo con una posición favorable en la medida que el nodo está situado entre los caminos entre otros pares de actores en la red.

$$C_B(i) = \sum_{j < k} g_{jk}(i) / g_{jk}$$

Pares de vértices posibles excluyendo el del mismo nodo

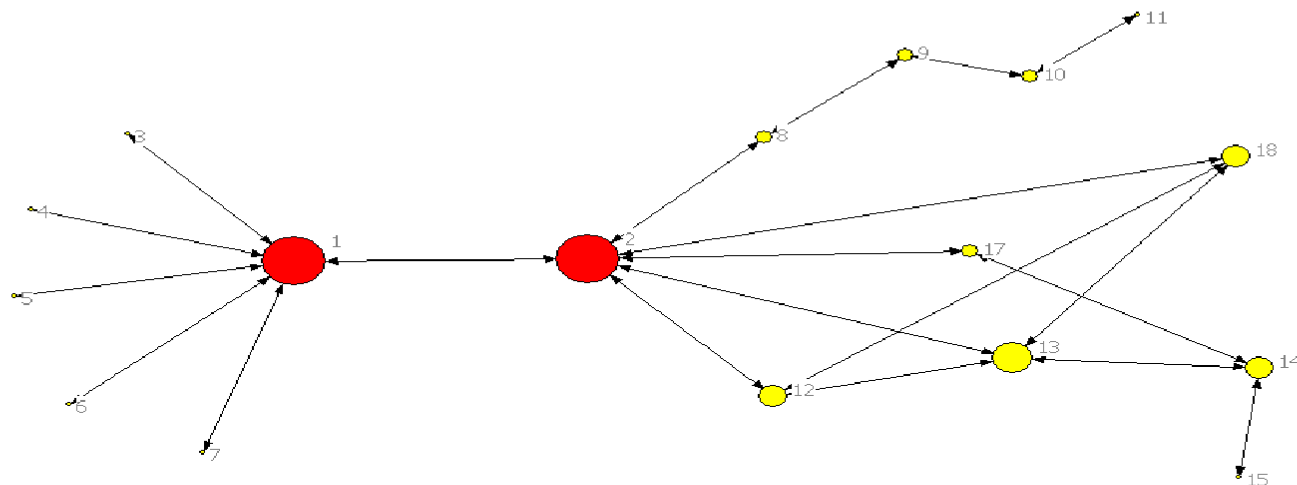
Donde g_{jk} = numero de caminos cortos que conectan j y k
 $g_{jk}(i)$ = numero de caminos cortos en los que el nodo i se encuentra.

$$C'_B(i) = C_B(i) / [(n-1)(n-2)/2]$$



MEDIDAS LOCALES DE CENTRALIDAD

Grado vs. Intermediación (1)

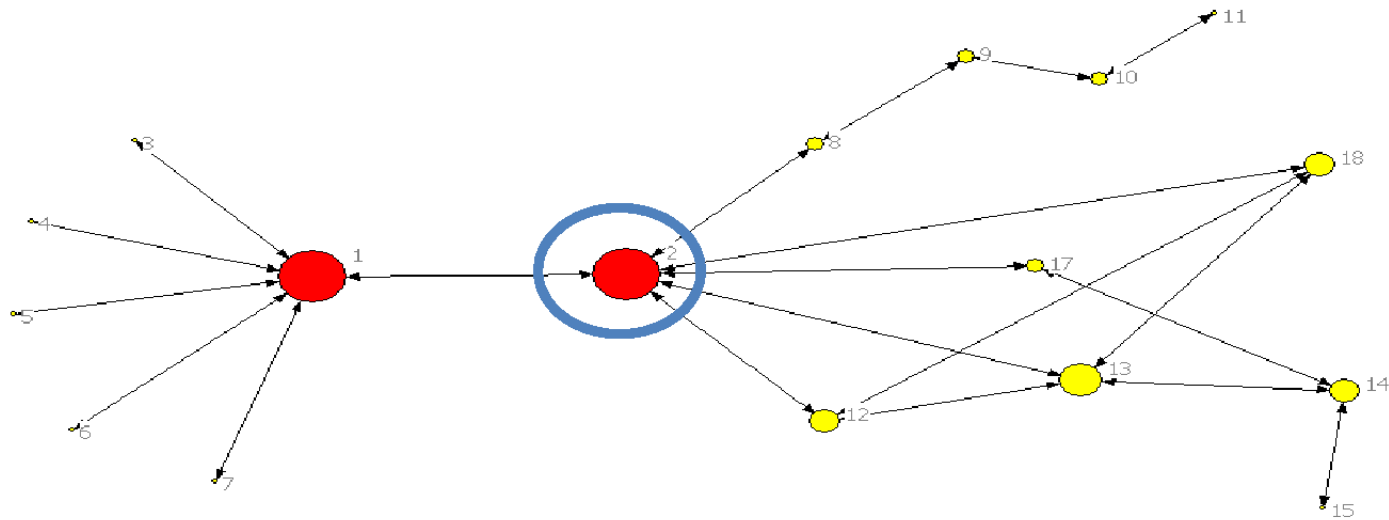


El grado de los nodos en rojo (1 y 2) es el mismo (6) pero, evidentemente, no son igual de importantes

¿Por qué? Por la debilidad de esta medida, que solo toma en cuenta los vínculos inmediatos (a un nivel local) dejando de lado los vínculos indirectos (a nivel global)

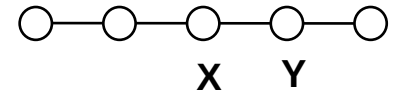
MEDIDAS LOCALES DE CENTRALIDAD

Grado vs. Intermediación (2)



Nombre	Grado (C_D) Max = 17	Intermediación (C_B) Max = $17 \cdot 16 / 2 = 136$
1	6	70.00
2	6	96.50

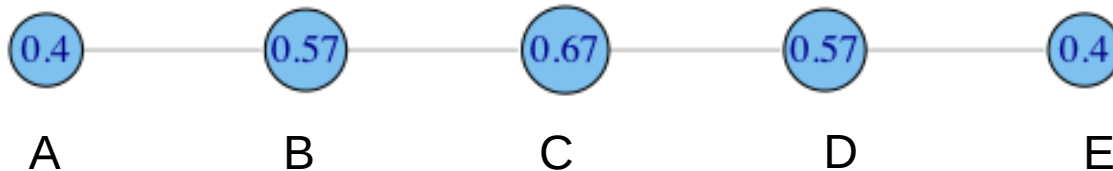
Métricas: closeness



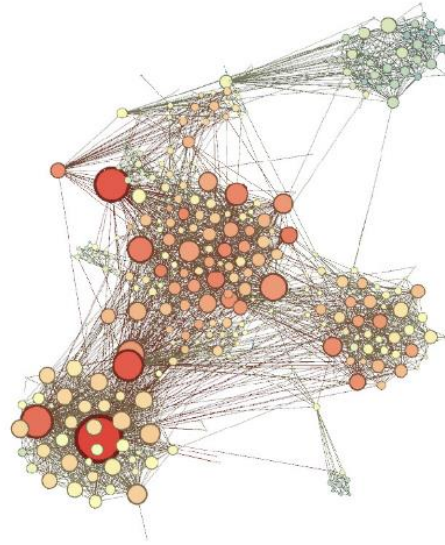
Distancia promedio del camino mas corto entre un nodo a todos los nodos.

Closeness Centrality: $C_c(i) = \frac{1}{N-1} \sum_{j=1}^N d(i, j)$

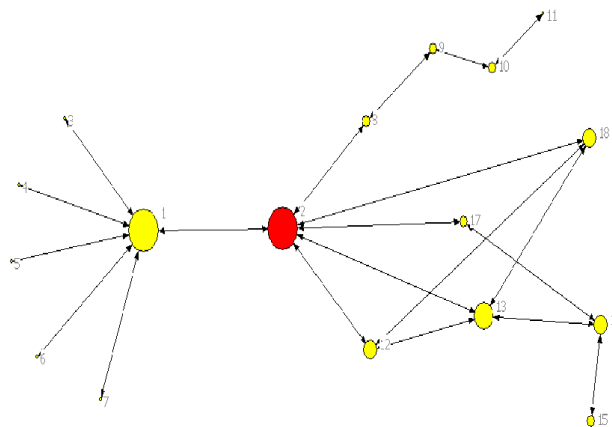
Normalized Closeness Centrality $C'_c(i) = (C_c(i)) / (N - 1)$



$$C'_c(A) = \frac{1}{N-1} \sum_{j=1}^N d(A, j) = \frac{1+2+3+4}{4} = \frac{10}{4} = 0.4$$



Red Personal de Contactos de Facebook de Oscar Cordón: el tamaño de los nodos indica el **grado** y el color la **cercanía** (más **azul**, menor valor; más **rojo**, mayor valor)



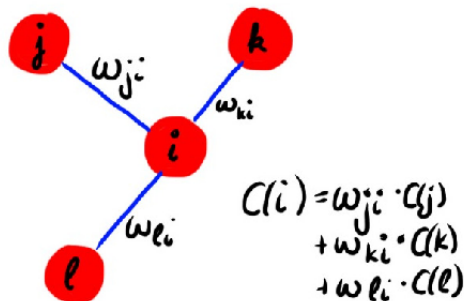
Actor	Lejanía	CercaníaN
2	34.000	50.000
1	40.000	42.500
13	42.000	40.476
17	44.000	38.636
8	44.000	38.636
12	45.000	37.778
18	45.000	37.778
14	52.000	32.692
6	56.000	30.357
5	56.000	30.357
7	56.000	30.357
3	56.000	30.357
4	56.000	30.357
9	56.000	30.357
15	66.000	25.758
10	70.000	24.286
16	82.000	20.732
11	86.000	19.767

MEDIDAS LOCALES DE CENTRALIDAD

Centralidad de vector propio (1)

La **Centralidad de vector propio** se basa en que la centralidad de un nodo concreto depende de cómo de centrales sean sus vecinos (**prominencia**)

La idea básica es que el poder y el status de un actor (**ego**) se define recursivamente a partir del poder y el status de sus vecinos (**alters**)



w_{ij} (a_{ij}) corresponde a la entrada de la matriz de adyacencia. Puede ser binaria $\{0,1\}$ o un peso numérico

La medida es válida para redes dirigidas (**Prestigio de rango**) y no dirigidas

Es una versión más elaborada de la Centralidad de grado al asumir que no todas las conexiones tienen la misma importancia. No se tiene en cuenta la cantidad sino la calidad de las mismas

Métricas: Eigenvector centrality Bonacich

$$c_i(b) = \frac{1}{\alpha} (a + \sum_j b c_j) A_{ji}$$

$$c(b) = a(I - bA)^{-1} A \mathbf{1}$$

- α constante de normalización
- β determina importancia de la centralidad de los vecinos
- A matriz de adjacencia (puede tener pesos)
- I matriz de identidad
- $\mathbf{1}$ matriz de unos.

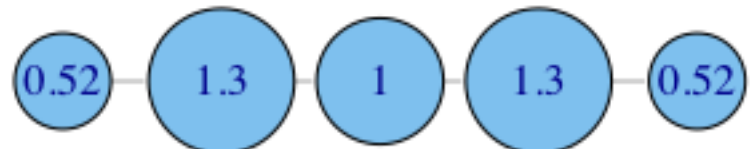
Si $\beta > 0$, los nodos tienen mayor centralidad cuando tienen enlaces con nodos con alta centralidad

$\beta = .25$



Si $\beta < 0$, los nodos tienen mayor centralidad cuando tienen enlaces a nodos no tan centrales

$\beta = -.25$



MEDIDAS LOCALES DE CENTRALIDAD

Excentricidad

Otra medida local de centralidad basada en distancias es la **Centralidad de excentricidad (C_E)**. Se define como la inversa de la **excentricidad** (la máxima distancia geodésica) entre un actor y cualquier otro actor de la red:

$$C_E(i) = \frac{1}{\max_{j \in V(G)/i} d(i, j)}$$

$$C'_E(i) = C_E(i) / g - 1$$

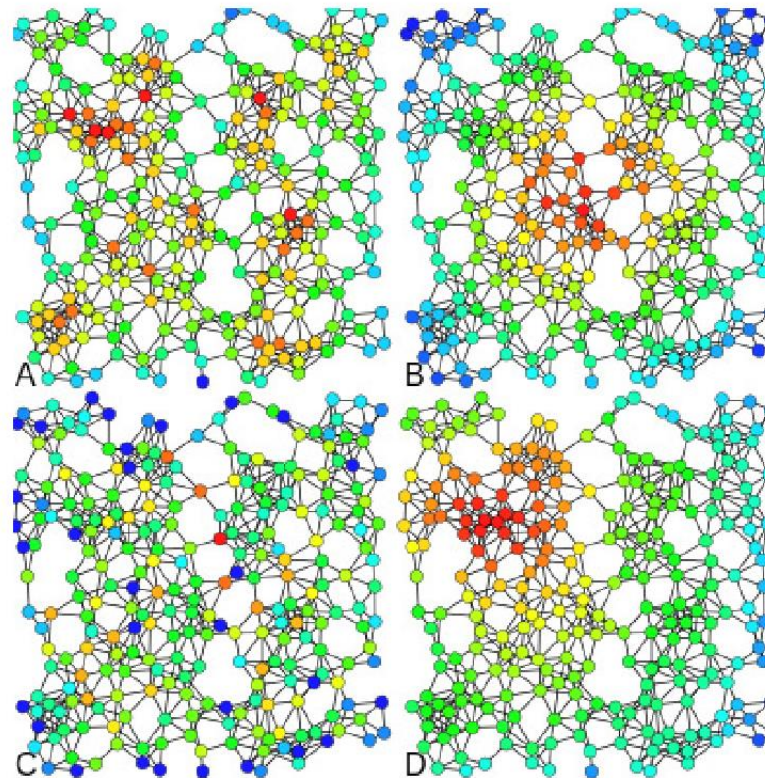
Los actores con un mayor valor de excentricidad se denominan **actores periféricos**, los de menor valor forman el **centro de la red**

MEDIDAS LOCALES DE CENTRALIDAD

Comparativa

A) centralidad de grado; B) cercanía; C) intermediación; D) centralidad de vector propio

azul = menor valor

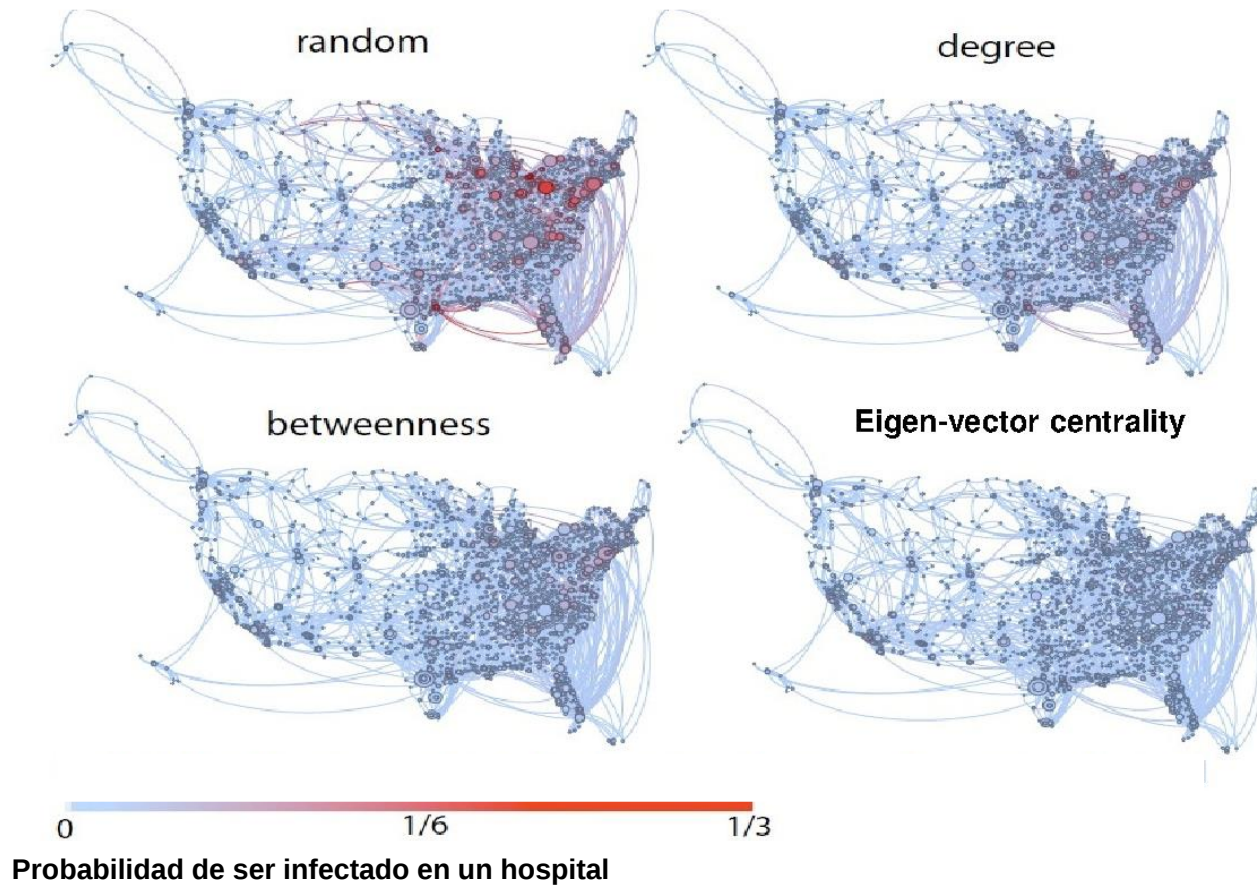


rojo = mayor valor

APLICACIONES

Infecciones en hospitales EEUU por transferencia de pacientes (4)

Red de transferencias de pacientes hospitalarios
Presupuesto para estrategias para evitar propagación de infecciones

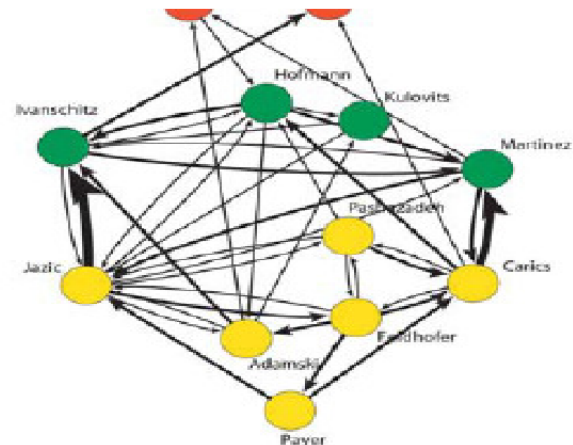
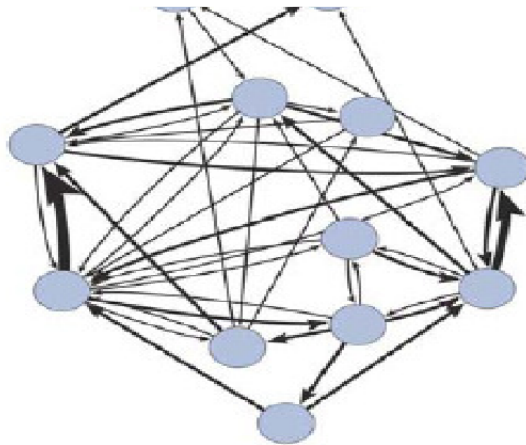


APLICACIONES

Análisis de juego en equipos de fútbol (1)

Descripción de un equipo de futbol de Viena llamado Rapid durante un juego realizado en diciembre del 2003

Describe el comportamiento en el Segundo tiempo



Grafo que describe los pases entre los jugadores

APLICACIONES

Análisis de juego en equipos de fútbol (2)

Posibles factores de análisis:

- ¿Qué jugador ha iniciado más pases (grado ponderado de salida)? **Jazic**
- ¿Qué jugador ha recibido más pases (grado ponderado de entrada)? **Jazic**
- ¿Quién ha controlado el juego del Rapid (centralidad)? **Jazic** y **Hoffman**
- ¿Qué jugadores han estado implicados en jugadas con el mayor número de pases (camino)? **Jazic**, **Hofmann**, **Feldhofer**, **Martinez** y **Carics**
- ¿Quién ha jugado con quién y quién no (análisis de los enlaces)? **Ni un solo pase de Ivanschitz a Wagner**
- ¿Qué grupos de jugadores han compuesto la columna vertebral del equipo (análisis de triadas)? **Por ejemplo, Feldhofer-Carics-Pashazadeh**
- ¿Qué jugadores han tenido un rol similar (análisis de enlaces)? **Por ejemplo, Ivanschitz / Martinez**



APLICACIONES

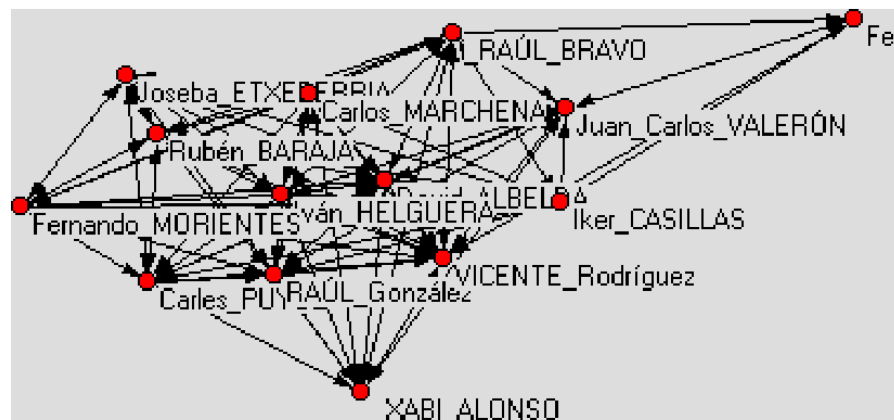
Análisis de juego en equipos de fútbol (4)

J.J. Merelo. Redes contra redes: el fútbol es así. <http://atalaya.blogalia.com/historias/19642>

Euro-copa Portugal 2004

Extraer las estadística de pases

España 1 – Rusia 0



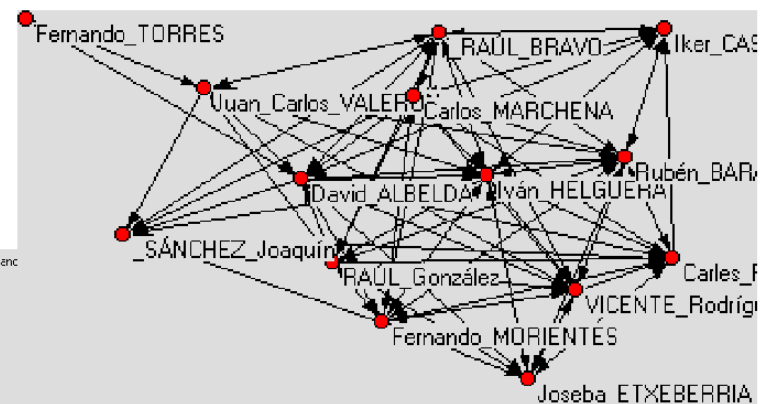
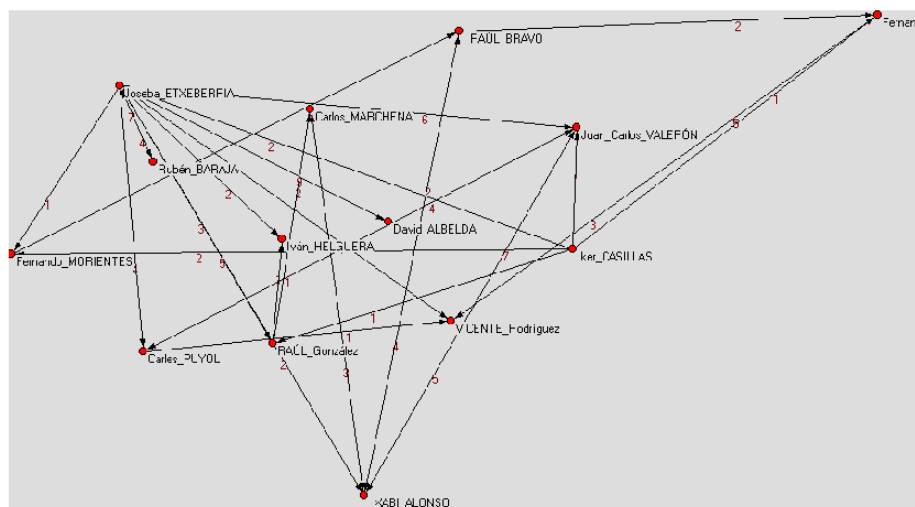
Una red donde, curiosamente, el jugador que tiene más centralidad es Iker Casillas, cuando debería ser un medio como Baraja

APLICACIONES

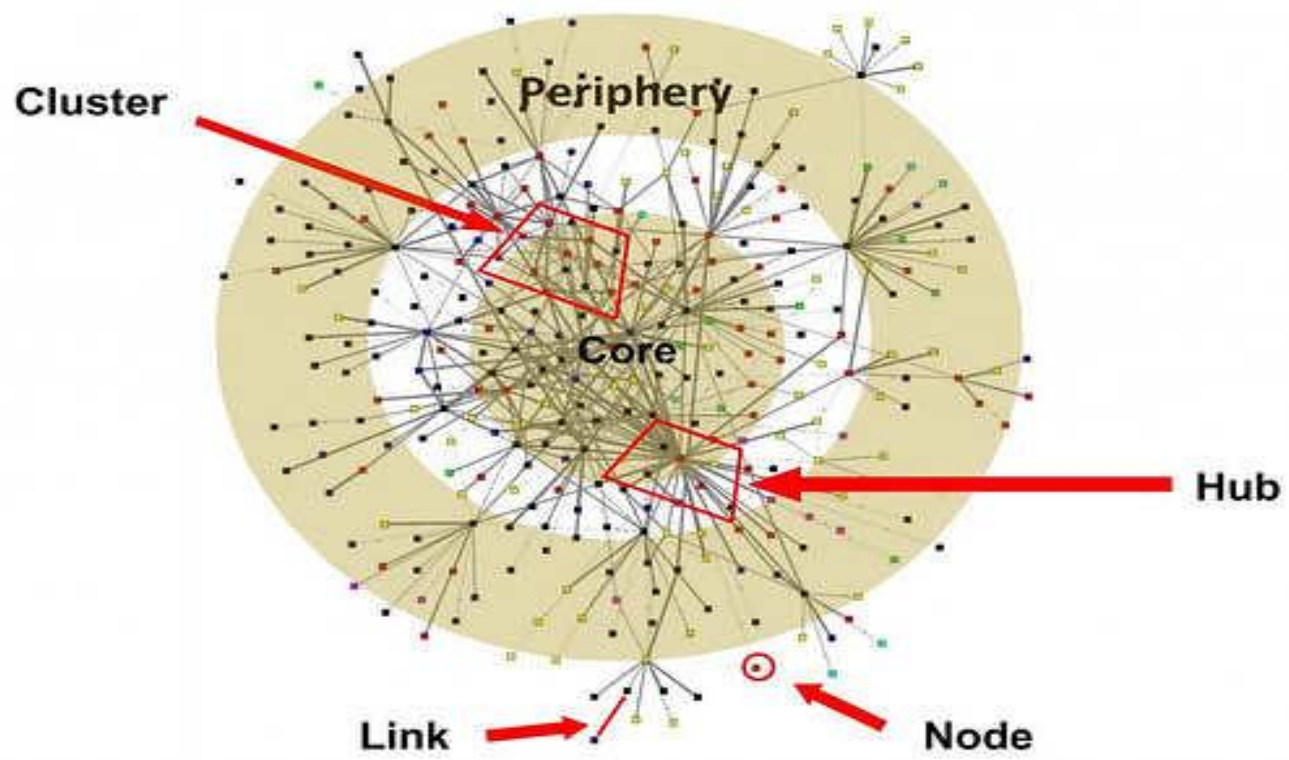
Análisis de juego en equipos de fútbol (5)

La situación no cambió mucho en el segundo partido (Grecia 1 – España 1) salvo que, en este caso, Albelda, Baraja y Helguera organizaron un poco más el juego. Y Fernando Torres a su bola, claro

Casi el 90% de los pases fueron los mismos. Esta es la diferencia de las dos redes (sin Joaquín porque es un nodo nuevo)



Es curioso ver también que la "autoridad" de la red es Vicente, un extremo. Lo lógico sería que las autoridades fueran los delanteros, pero Morientes y Raúl se hallan ahí perdidos, en la maraña de la red



Métricas: Comunidades

☐ Mutualidad

- ☐ Cada miembro conoce a todos los miembros

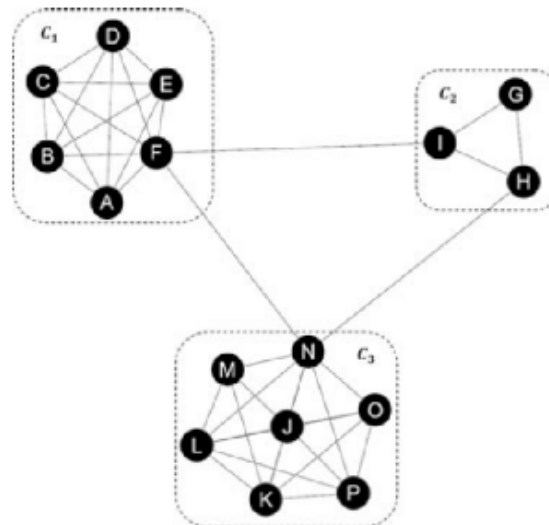
☐ Frecuencia

- ☐ Cada miembro conoce al menos k miembros del grupo

☐ Cercanía

- ☐ Los miembros están separados por máximo de n saltos

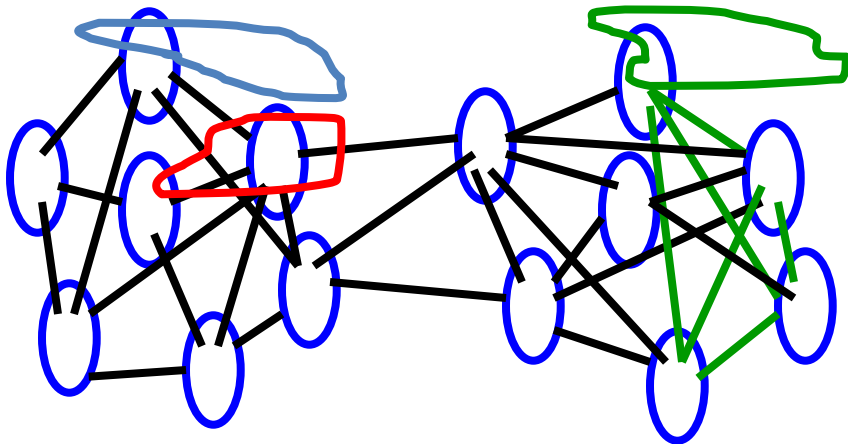
Métricas: Comunidades



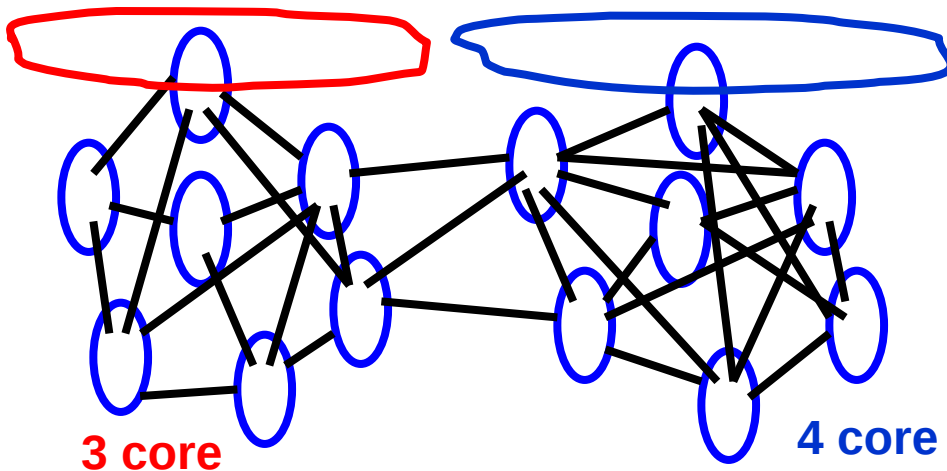
- En esta red, hay **tres comunidades**: C_1 , C_2 y C_3
- Cada comunidad está formada por un grafo completo (un **clique**) de tamaño variable ($C_1 = K_6$, $C_2 = K_3$ y $C_3 = K_7$)
- La densidad de enlaces entre las comunidades es muy baja. Los pocos enlaces que existen son **puentes**

Cliques y K-core

Cada miembro del grupo posee un link a todos los miembros del grupo



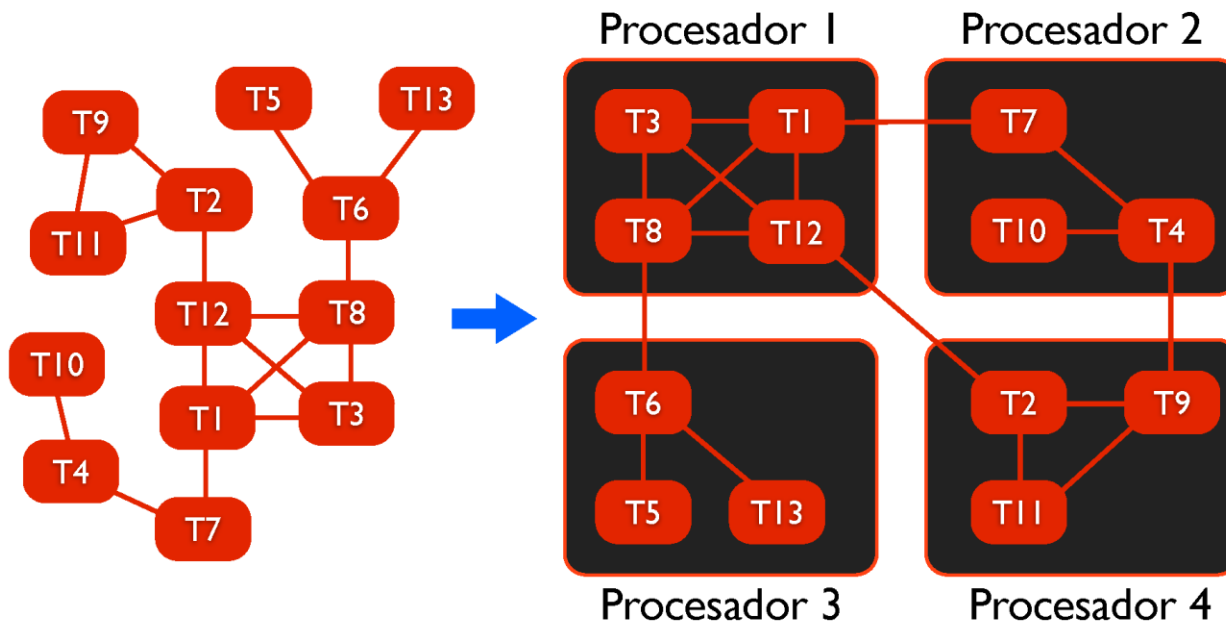
- ❑ Si se pierde un enlace, deja de ser un clique
- ❑ No es interesante que todos estén conectados con todos
- ❑ No hay medidas de centralidad dentro de un clique



Cada miembro del grupo está conectado con k otros miembros del grupo

Computación en Paralelo

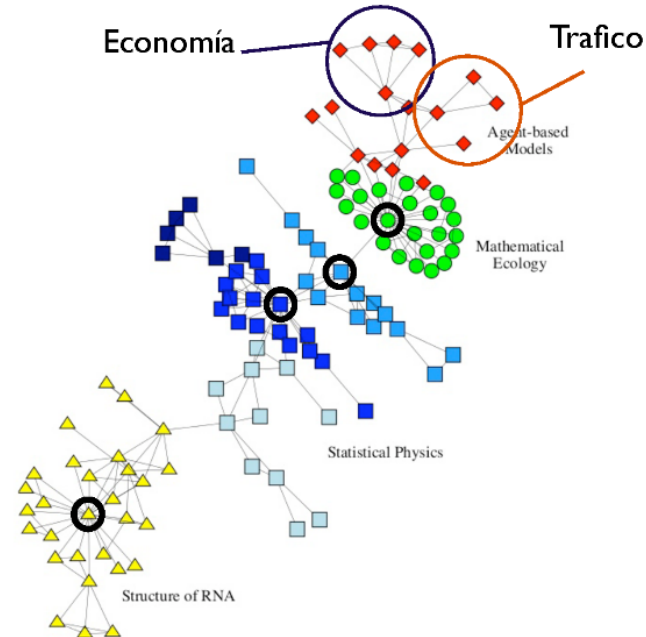
(Particionado de Grafos)



Distribución mas eficiente de tareas en un conjunto de procesadores

Redes de colaboración científica

- ◆ Modelos basados en agentes para estudiar problemas de economía y flujo de tráfico
 - Modelos matemáticos en ecología
 - Física estadística
 - ▲ Estructura del ARN
- Formación de comunidades en torno a la metodología
 - El centro de las comunidades corresponde al jefe del grupo de investigación.



Red de colaboración de científicos del Instituto de Santa Fe en Nuevo México

M. Girvan and M. E. J. Newman (2002). "Community structure in social and biological networks". Proc. Natl. Acad. Sci. USA 99 (12): 7821–7826. doi:10.1073/pnas.122653799. PMC 122977. PMID 12060727.

Métricas: Comunidades

Algoritmos de detección

- Edge Betweenness Method, M. Girvan and M. E. Newman (GN)
- Fast greedy modularity optimization, A. Clauset, M. E. Newman, and C. Moore (Clauset et al.)
- Exhaustive modularity optimization via simulated annealing, R. Guimerá, M. Sales-Pardo (Sim ann.)
- Multi-Level Aggregation Method based on modularity, V. D. Blondel, J.-L. Guillaume, R. Lambiotte (Blondel et al.)
- Divisive algorithm based on the edge-clustering coefficient, F. Radicchi, C. Castellano, F. Cecconi (Radicchi et al.)
- Clique Percolation Method for finding communities, G. Palla, I. Derenyi, I. Farkas (Cfinder)
- Graph clustering by flow simulation, S. van Dongen (MCL)

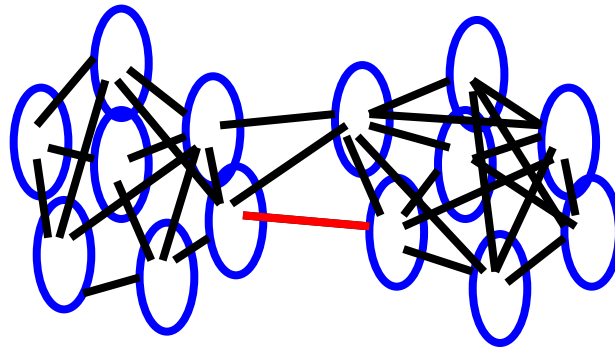
Betweenness clustering

Entrada: Grafo inicial $G(E,V)$

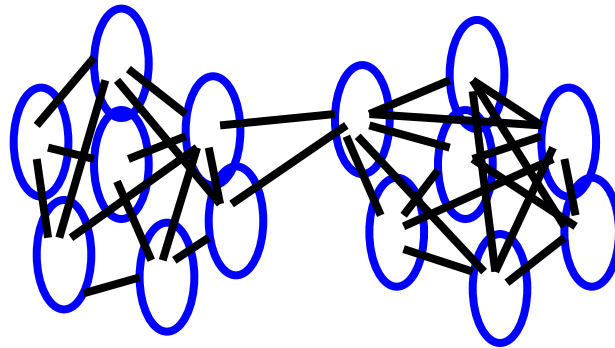
Inicio:

1. Calcular la intermediación para todas las aristas en la red.
2. Generar un nuevo grafo al remover las aristas con un valor alto de intermediación.
3. Recalcular la intermediación para todas las aristas del nuevo grafo.
4. Repetir el paso 2 hasta que todas las aristas tengan un nivel bajo de intermediación.

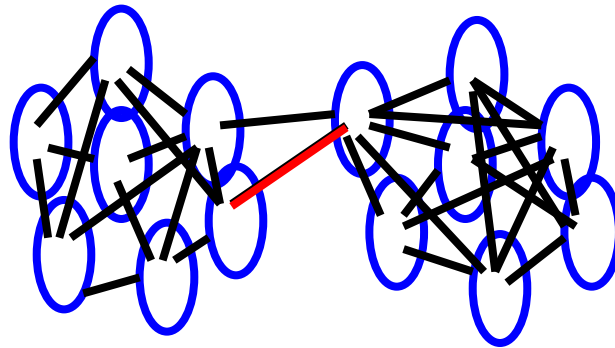
Betweenness clustering



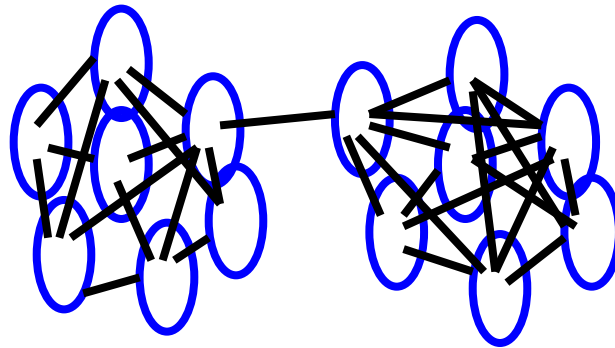
Betweenness clustering



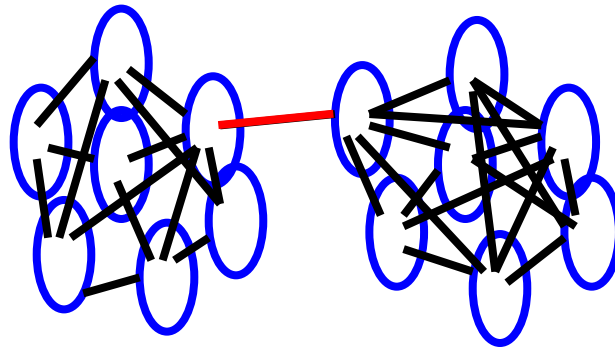
Betweenness clustering



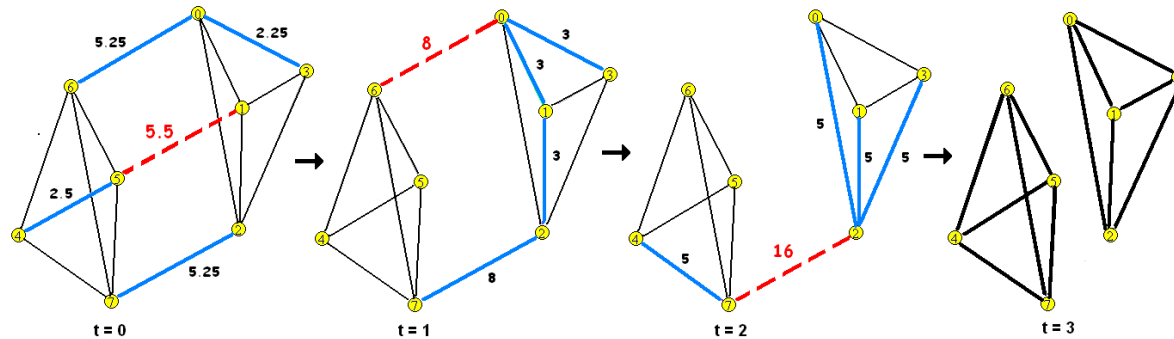
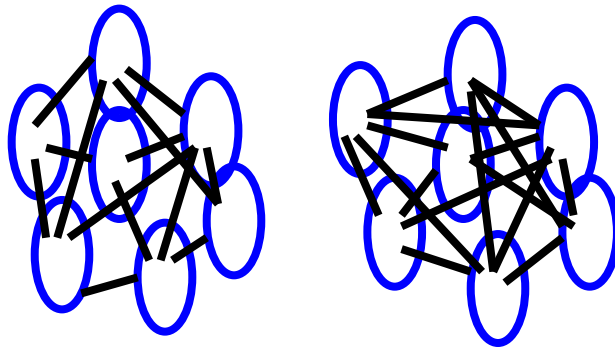
Betweenness clustering



Betweenness clustering



Betweenness clustering



Modularidad

- Métrica diseñada para dividir la red en módulos, clusters, comunidades.
- Es usada para maximizar métodos para hallar comunidades
- Una red con alta modularidad significa que:
 - Es muy densa entre nodos de una misma comunidad
 - Pero con conexiones dispersas entre nodos de comunidades distintas

Modularidad

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

1 Si v y w están en la misma comunidad, 0 si no

Matriz de adyacencia

Probabilidad de un arco entre dos nodos, es proporcional a sus grados

Detección basada en la Modularidad

Heurísticas utilizadas para la optimización de la modularidad:

- Recocido Simulado (Guimera and Amaral)
- Optimización extrema (J. Duch and A. Arenas)
- Algoritmos voraces (Clauset et al.)
- Reformulación de la modularidad en términos de las propiedades espectrales de la red. (Newman)

Las dos ultimas heurísticas han resultado efectivas sin embargo poseen un problema inherente al concepto de modularidad llamado *limite de resolución*.

Modularity and community structure in networks, M. E. J. Newman, Proc. Natl. Acad. Sci. USA 103, 8577–8582 (2006).

Wikipedia, 2011. Modularity (networks). [en línea] Disponible en: <[http://en.wikipedia.org/wiki/Modularity_\(networks\)](http://en.wikipedia.org/wiki/Modularity_(networks))>
[Consultado el 17 Noviembre 2011].

MEDIDAS GLOBALES

Existen varias medidas globales en SNA. La mayoría son las mismas empleadas para analizar cualquier otro tipo de red, que ya hemos estudiado:

1. **Diámetro y Radio**
2. **Distancia media**
3. **Grado Medio**
4. **Densidad**
5. **Coeficiente de Clustering Global**
6. **Reciprocidad**

Diámetro ($d_{\max}$): longitud del camino mínimo más largo de la red

En redes grandes, se puede determinar con el algoritmo de búsqueda primero en anchura

Equivale al valor máximo de excentricidad para todos los nodos de la red:

$$E(i) = \max_{j \in V(G) / i} d(i, j) \quad d_{\max} = \max \{E(i) : i \in V(G)\}$$

En el contexto del SNA, esta métrica da una **idea de la proximidad entre pares de actores en la red**, indicando cómo de lejos están en el peor de los casos

Las redes más dispersas suelen tener un mayor diámetro que las más densas al existir menos caminos entre cada par de nodos

Radio (r): Valor mínimo de excentricidad para toda la red:

$$r = \min \{E(i) : i \in V(G)\}$$

► **Densidad:** Actividad global de la red

► En redes no dirigidas:

$$\Delta = L / [g(g-1)/2]$$

► En redes dirigidas:

$$\Delta = L / g(g-1) \quad \begin{array}{l} L, \text{ número de enlaces} \\ g, \text{ número de nodos} \end{array}$$

Distancia media ($\langle d \rangle$) para un grafo dirigido:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i, j \neq i} d_{ij} \quad d_{ij} \text{ es la distancia geodésica entre los nodos } i \text{ y } j$$

En un **grafo no dirigido** $d_{ij} = d_{ji}$. De este modo, sólo es necesario contar la longitud de los caminos una vez:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i, j > i} d_{ij}$$

La medida da una idea de cómo de lejos están los distintos actores en promedio. En SNA representa la **eficiencia del flujo de información en la red**

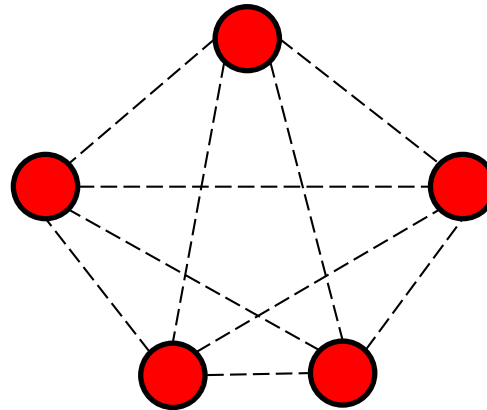
Medida de red : Densidad de una red no dirigida

En el siguiente grafo:

Las conexiones posibles son:

10 conexiones posibles en el grafo.

Pos.= posibles
Eff . = efectivas



Medidas de red
Densidad de una red no dirigida

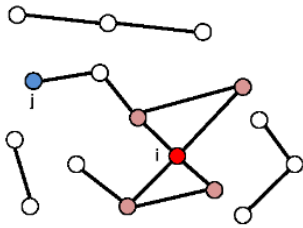
$$d = \frac{n^{\circ} \text{ effective edges}}{n^{\circ} \text{ possible edges}}$$

Eff.=0 Pos.=10	Eff.=2 Pos.=10	Eff.=4 Pos.=10	Eff.=8 Pos.=10	Eff.=10 Pos.=10
d=0	d=0.2	d=0.4	d=0.8	d=1

MEDIDAS GLOBALES

Grado medio

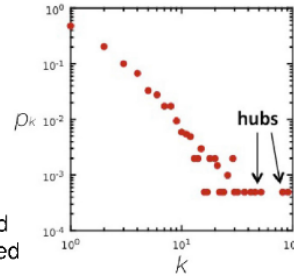
No dirigida



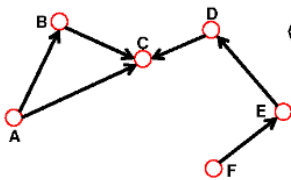
$$\langle k \rangle \equiv \frac{1}{N} \sum_{i=1}^N k_i$$

$$\langle k \rangle \equiv \frac{2L}{N}$$

N – número de nodos de la red
L – número de enlaces de la red



Dirigida



$$\langle k^{in} \rangle \equiv \frac{1}{N} \sum_{i=1}^N k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N} \sum_{i=1}^N k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

$$\langle k \rangle \equiv \frac{L}{N}$$

MEDIDAS GLOBALES

Densidad y Coeficiente de Clustering

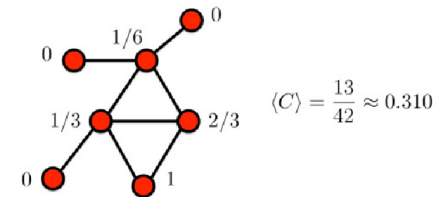
La **densidad D** = L/L_{max} mide el grado de conectividad de la red social a nivel global

El coeficiente de clustering C_i mide la **densidad local**, el ratio de vecinos de un nodo que están conectados entre sí

El **coeficiente de clustering medio** ($\langle C \rangle$) es una **medida global** que indica la probabilidad de que dos vecinos de un nodo de la red escogido aleatoriamente estén conectados entre sí:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i$$



Las redes sociales reales son redes de mundos pequeños (*small worlds*) y suelen tener valores de $\langle C \rangle$ altos. Eso implica que la **transitividad** entre nodos aparece más y con más fuerza, incrementando la probabilidad de que se formen **cliques**

MEDIDAS GLOBALES

Reciprocidad

La **reciprocidad (R)** en un **grafo dirigido** mide la tendencia de pares de actores a tener conexiones mutuas entre ellos

La forma más habitual de calcularla es como un ratio entre el número de conexiones (*díadas*) mutuas (*#mut*) y el número de conexiones totales en la red, las mutuas y las asimétricas (*#asim*):

$$R = \frac{\#mut}{\#mut + \#asim}, \quad R \in [0,1]$$

Indica la probabilidad de que dos actores de la red apunten el uno al otro

Una *díada asimétrica* es un par de actores que presentan un arco en una u otra dirección pero no en ambas. Una *díada mutua* incluye los arcos en las dos direcciones

APLICACIONES

Aprendizaje Colaborativo por Ordenador (1)

Aprendizaje Colaborativo por Ordenador: énfasis en las relaciones entre los actores en un curso on-line en BSCW

Relación entre el profesor y los alumnos en un curso, así como entre los propios alumnos de distintos grupos

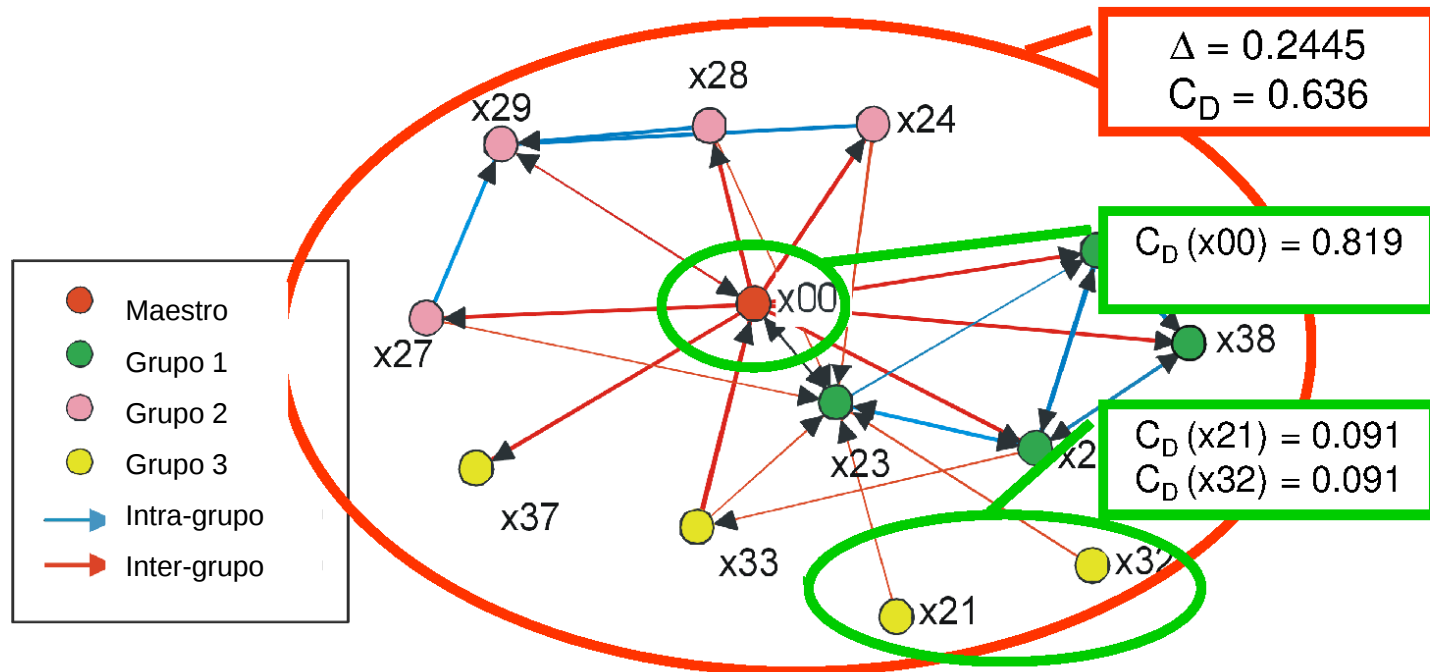
Red Social con distintas variantes: trimodal egocéntrica (tres grupos de alumnos-profesor) y trimodal, unimodal completa (todos los alumnos con todos) y trimodal completa (miembros de un grupo con los de otros)

Pregunta global: ¿Cómo ayudar a los profesores a monitorizar aspectos colaborativos de aprendizaje mediante la tecnología?

Análisis con dos medidas globales (Densidad de la red Δ y Centralización de grado C_D) y dos medidas locales (Centralidades de grado y de cercanía)

APLICACIONES

Aprendizaje Colaborativo por Ordenador (2)

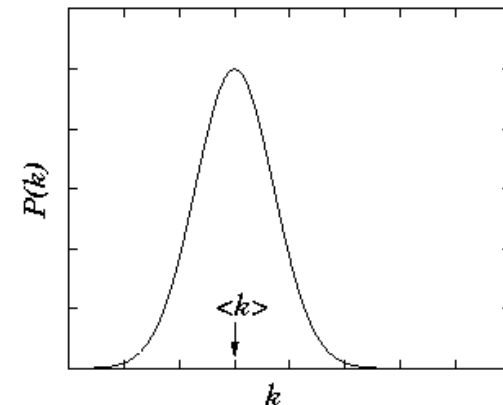


Grafos aleatorios

En Matemáticas se denomina **grafo aleatorio** a un **grafo** que es **generado** por algún **tipo de proceso aleatorio**.

- Nodos conectados al azar
- Número de aristas incidentes en cada nodo se distribuye Poisson

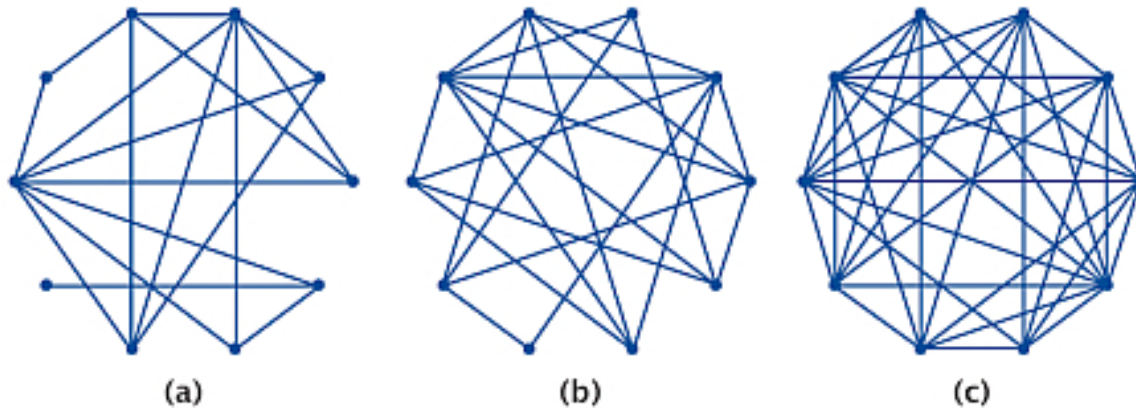
Distribución de Poisson



Grafos Aleatorios

Iniciador de la Teoría de Redes Complejas: Erdos-Renyi (50's).

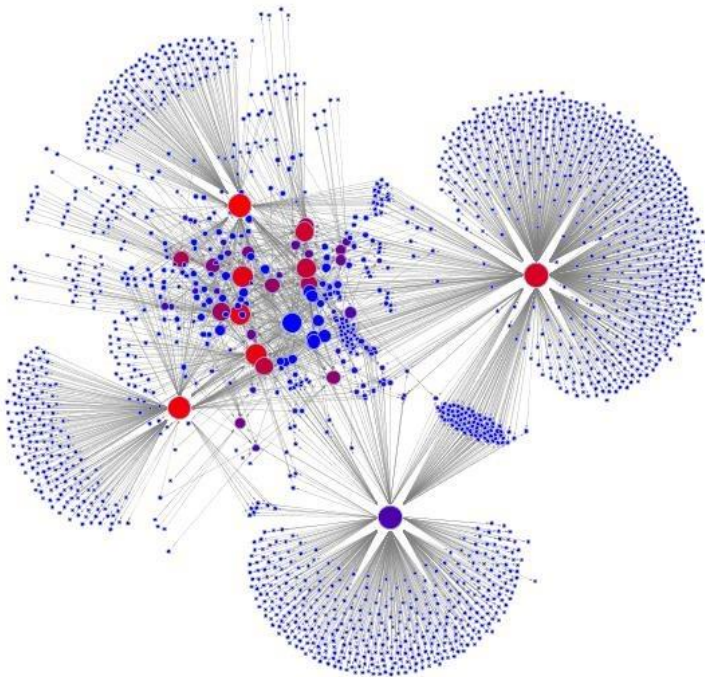
- Construye la red enlazando nodos elegidos al azar según determinada probabilidad



Grafos Aleatorios con distinta probabilidad de conexión entre pares de nodos (A) $p=0.3$, (B) $p=0.5$ y (C) $p=0.8$

En este **modelo Erdős–Rényi** se tiene que **un nuevo nodo se enlaza con igual probabilidad con el resto de la red,**

Power-law



- Nodos aparecen con el tiempo (growth model)
- Nodos prefieren unirse a nodos populares (preferential model)
- Nodos viejos Mueren
- Algunos nodos son mas sociable
- Las amistades pueden desaparecer

Propagación de epidemias

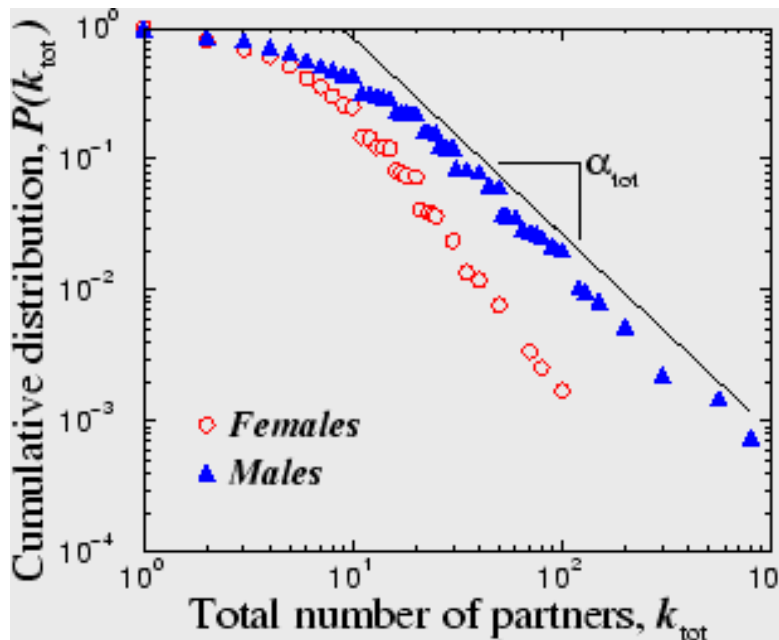
El papel de los centros en las epidemias

- En una red del tipo power-law, un virus puede persistir por más baja sea su capacidad de infección
- Muchas redes del mundo real hacen exhibir power-laws:
 - contactos sexuales
 - redes de correo electrónico

Propagación de epidemias: **RED DE CONTACTOS SEXUALES**

Nodos: individuos

Links: relaciones sexuales



HAY UNOS POCOS
NODOS CON MAYOR
PROBABILIDAD DE
CONTAGIAR QUE OTROS
(HUBS)



ESTRATEGIAS DE
PREVENCION DE EPIDEMIAS

Difusión de información en las redes

- **Factores que influyen en la difusión de información**
 - estructura de la red: la que se conectan los nodos?
 - fuerza de los lazos: qué tan fuerte son las conexiones?
- **Estudios en la difusión de la información:**
 - Granovetter: la fuerza de los lazos débiles
 - J-P Onnela et al: fuerza de los lazos intermedios
 - Kossinets et al: fuerza de los lazos de backbone
 - Davis: enclavamientos de mesa y adopción de practices

Aplicaciones

- Epidemiología (propagación de virus).
- Tolerancia frente a ataques (deliberados).
- Procesos de optimización (publicidad).
- . . .

Minería de Grafos

Objetivo: Desarrollar algoritmos para extraer y analizar grafos.

- Búsqueda de patrones en ellos
- Búsqueda de grupos de grafos similares (clustering)
- Construcción de modelos de predicción para los grafos (clasificación)
- Aplicaciones
 - descubrimiento de motivos estructurales
 - reconocimiento de proteínas
 - ingeniería inversa en VLSI
 - Mucho más ...

Por qué Minería de Grafos?

- **Los grafos son ubicuos**
 - Compuestos químicos (quimio-informática)
 - Estructuras de las proteínas, las vías/redes biológicas (Bioinformática)
 - Flujo de programas, flujo de tráfico, flujo de trabajo
 - bases de datos XML, Web, de redes sociales
- **Grafos es un modelo general**
 - Árboles, secuencias, lazos, etc.
- **Diversidad de grafos**
 - Dirigidos vs. no dirigidos, etiquetados vs. no etiquetados (arcos y vértices), ponderados, con ángulos y geometrías (topológicos en 2-D/3-D)
- **La complejidad de los algoritmos: muchos problemas son de alta complejidad**

Minería de Patrones de Grafos

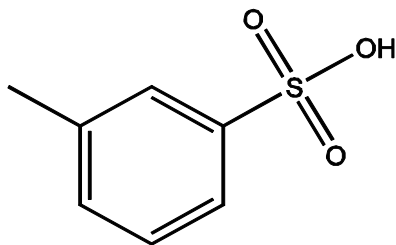
Minería subgrafo frecuentes

- **Encontrar subgrafos frecuentes dentro de un grafo**
 - SUBDUE (DOMINAR)
- **Encontrar (sub)grafos frecuentes en un conjunto de grafos**
 - *Support* (frecuencia de ocurrencia) no inferior a un umbral mínimo
 - Enfoques basado en Apriori
 - Enfoques de crecimiento del patrón (Pattern-growt)
- **Aplicaciones de la minería de patrones de grafos**
 - Minería de estructuras bioquímicas, de flujos de programas, de estructuras XML y comunidades de la Web
 - Construcción de sistemas de clasificación, agrupación, compresión, comparación y análisis de correlación para grafos

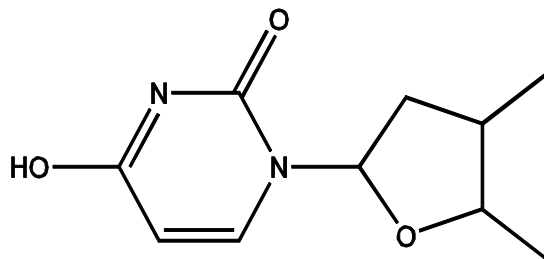
Enfoques de Minería subgrafo frecuentes

- **Enfoques basados en Apriori**
 - AGM: Inokuchi, et al. (PKDD'00)
 - FSG: Kuramochi and Karypis (ICDM'01)
 - PATH: Vanetik and Gudes (ICDM'02, ICDM'04)
 - FFSM: Huan, et al. (ICDM'03)
- **Enfoques de crecimiento del patrón (Pattern-growth)**
 - MoFa, Borgelt and Berthold (ICDM'02)
 - gSpan: Yan and Han (ICDM'02)
 - Gaston: Nijssen and Kok (KDD'04)
- **Minería de patrones cercanos**
 - CLOSEGRAPH: Yan & Han (KDD'03)

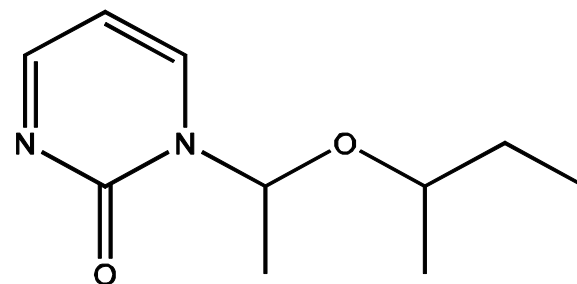
Descubrir subgrafos repetitivos (patrones) en Compuestos químicos



(A)



(B)



(C)

El soporte de un patrón dado, P, es :

$$sup_P = \sum_{G' \in D_P} S(P, G') / |D|$$

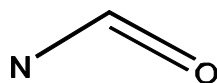
D colección de grafos

D_P representa el subconjunto de grafos en D que contienen un subgrafo aproximadamente isomorfo a P, dado el umbral de similitud τ .

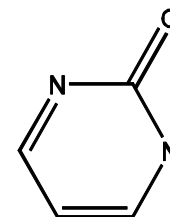
Un patrón P es frecuente si su soporte es mayor o igual que un umbral de frecuencia σ dado por el usuario.

FRECUENTES PATRONES (MIN SOPORTE ES >2)

(1)

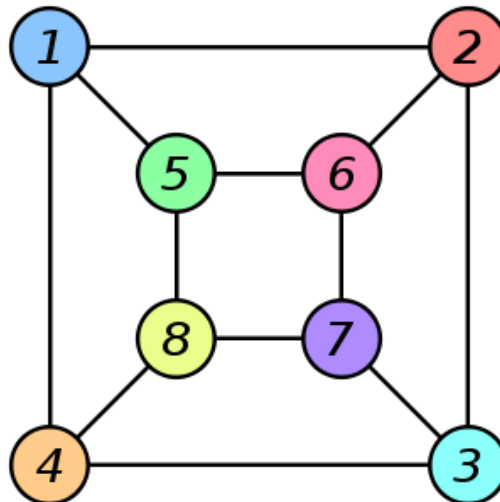
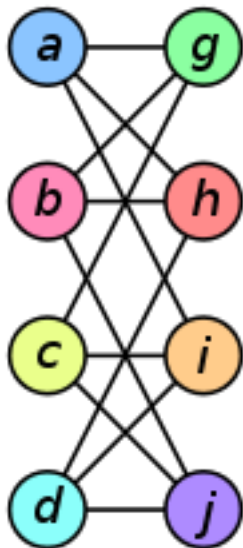


(2)



Descubrir subgrafos repetitivos (patrones)

Isomorfismo entre dos grafos G y H es una biyección F entre los conjuntos de sus vértices $F: V(G) \rightarrow V(H)$ que preserva la relación de adyacencia.



$F(a)=1$
 $F(b)=6$
 $F(c)=8$
 $F(d)=3$
 $F(g)=5$
 $F(h)=2$
 $F(i)=4$
 $F(j)=7$

Medidas de similitud basadas en los patrones de grafos

- Medidas de similitud basado en características
 - Cada grafo se representa como un vector de características
 - Vector de distancia
- Medida de similitud basada en la Estructura
 - Subgrafo común máximo
 - Grafo edita distancia: inserción, supresión, y re-etiquetado
- Subgrafos frecuentes y discriminativos son características de indexación de alta calidad



www.ing.ula.ve/~aguilar
aguilar@ula.ve

Introducción a la Minería
Semántica



José Aguilar
Editor



“If you are looking for different results,
then do not always do the same”

A. Einstein