Reconstrução de Florestas para o Problema de Filogenia Multimídia

Filipe de Oliveira Costa

Orientador: Anderson Rocha

Coorientador: Zanoni Dias

Instituto de Computação – UNICAMP

29 de Novembro de 2013

REWIND
REVerse engineering of audio-VIsual coNtent Data

RECOD
reasoning for complex data

UNICAMP

Sumário

1. Introdução
2. Filogenia de Documentos
3. Proposta
4. Trabalhos Correlatos
5. Metodologia e Desafios
6. O que já foi feito?
7. Cronograma

Introdução

Introdução

- Aumento do número de usuários de Internet
- Facilidade de compartilhamento de conteúdo multimídia
    - Exemplos: YouTube e Flickr
- Consequências negativas
    - Quebra de direitos autorais
    - Compartilhamento de conteúdo ilegal

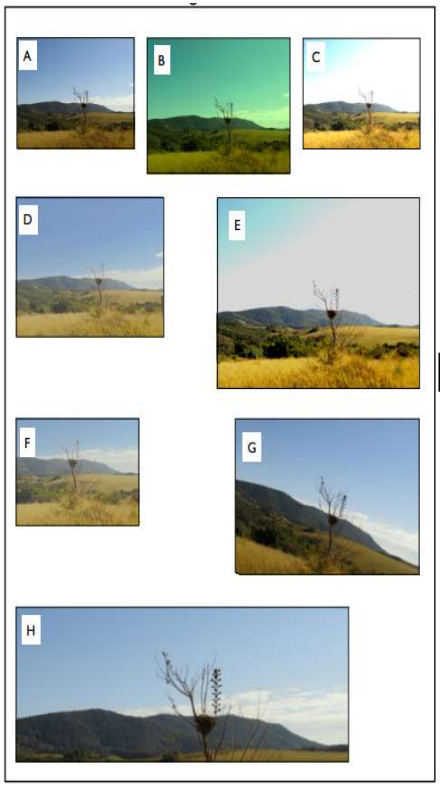
Introdução

- Alguns trabalhos tem como objetivo a detecção de duplicatas
  - Identificação de versões modificadas de um determinado documento
- Não avalia o histórico de geração das duplicatas

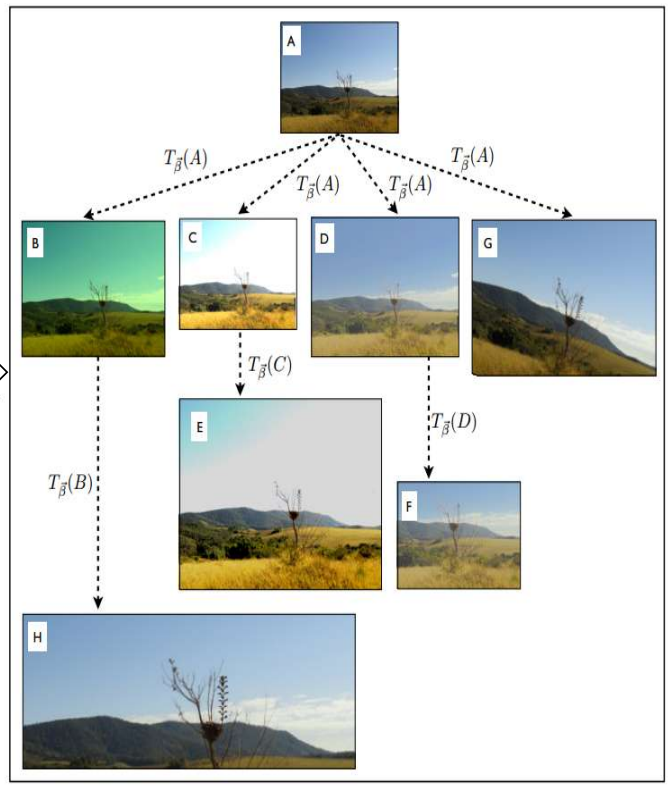
Filogenia de Documentos

Filogenia de Documentos

Conjunto de Duplicatas

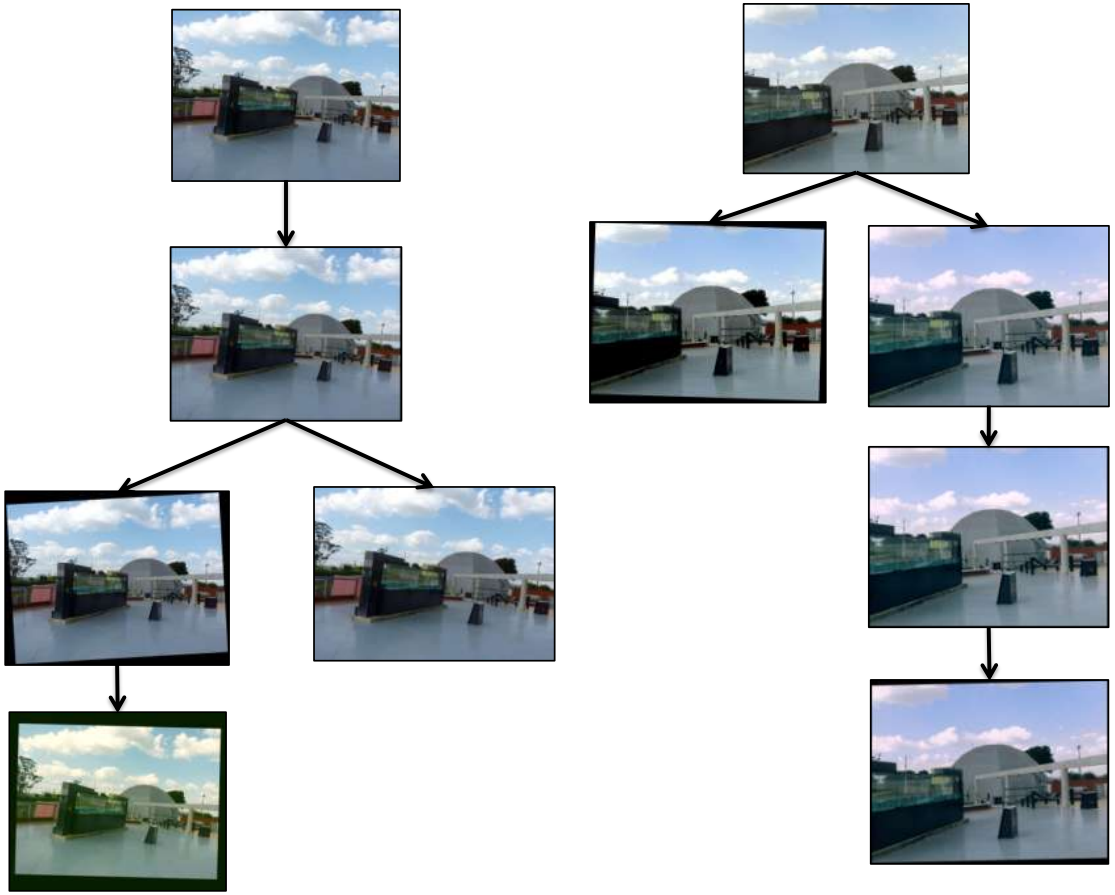


Árvore filogenética



Filogenia de Documentos

Filogenia de Documentos



Proposta

Proposta – Objetivo

- Desenvolver abordagens que permitam a identificação da estrutura de geração de um conjunto de duplicatas de documentos multimídia
- Consideramos como meta principal do projeto resolver o problema de encontrar florestas

Proposta – Aplicações

- Investigar quebra de direitos autorais
- Encontrar pistas sobre o criador de um dado documento com conteúdo ilegal
- Determinar o processo de formação de opinião no decorrer do tempo/espaço
- Melhor qualidade dos resultados na análise forense

Trabalhos Correlatos

Trabalhos Correlatos – Kennedy et al. (2008)[1]

- Kennedy et al. (2008) definiram a relação "pai-filho" entre as imagens
- Representação de uma aproximação do histórico da imagem
  - *Visual Migration Map*
- Poucos detalhes sobre:
  - Possíveis parâmetros para a família de transformações
  - Reconstrução da árvore de descendência

[1] L. Kennedy e S.-F. Chiang. "Internet Image Archaeology: Automatically tracing the manipulation history of photographs on the web". Em: *Proc. ACM Intl. Conference of Multimedia*. 2008, pp. 349–358.

Trabalhos Correlatos – De Rosa et al. (2010)[2]

- Abordagem para detectar a dependência de imagens
- Relação entre as imagens é descrita como a composição de dois componentes separadamente
  - Baseados no conteúdo da imagem
  - Independentes do conteúdo da imagem (PRNU)
- Reconstrução do grafo de dependência
  - Remoção de ciclos
  - Garantia que o grau de entrada de um vértice é igual a 1

[2] A. De Rosa et al. "Exploring image dependencies: a new Challenge in Image Forensics". Em: *SPIE-IS&T Electronic Imaging* 7541 (2010), pp. 774–788.

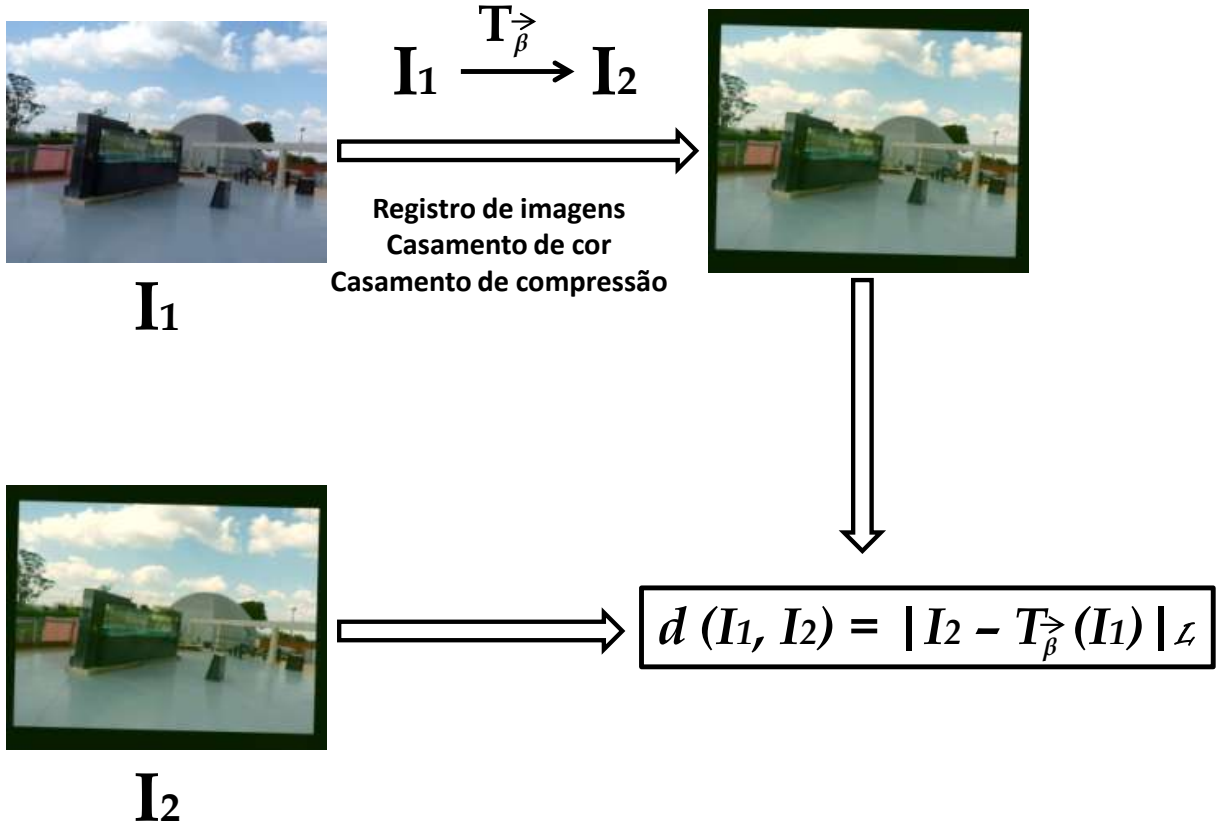
Trabalhos correlatos – Dias et al. (2010)[3]

- Dias et al. (2010) definiram formalmente o problema de filogenia de imagens
- Etapas
  - Cálculo da matriz de dissimilaridade
  - Algoritmo de reconstrução da árvore filogenética

[3] Z. Dias, A. Rocha e S. Goldenstein. "First steps towards image phylogeny". Em: *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*. 2010, pp. 1–6.

Trabalhos correlatos – Dias et al. (2010)

Cálculo da dissimilaridade

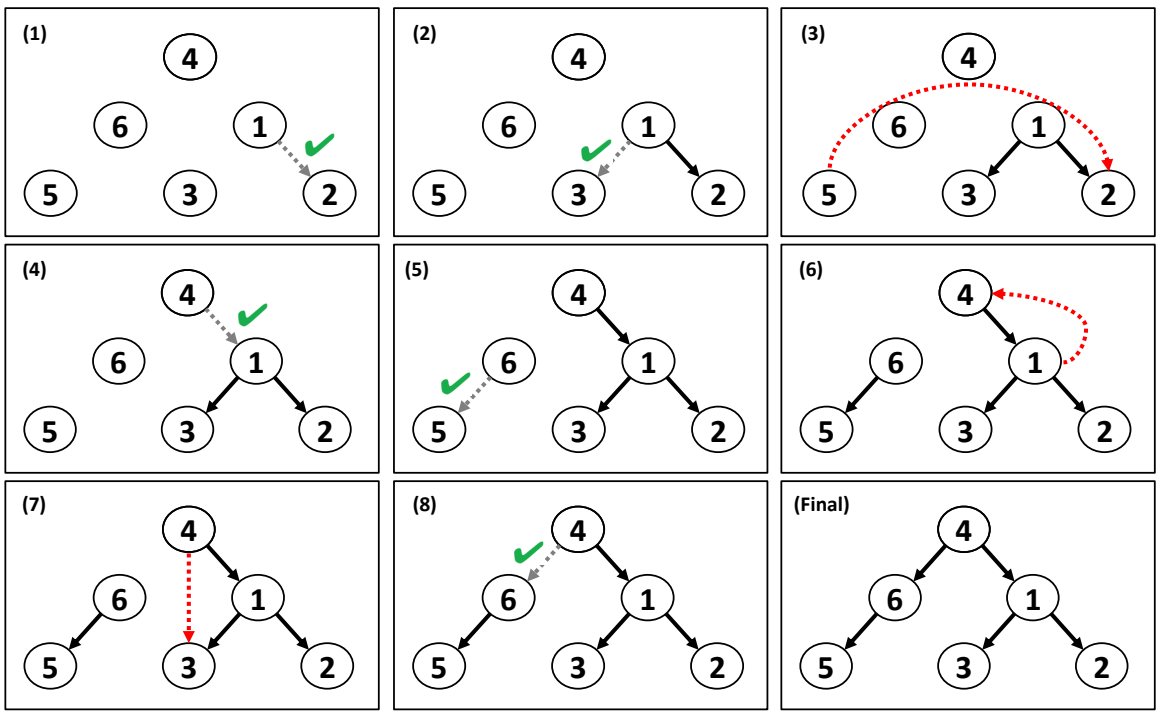


Trabalhos correlatos – Dias et al. (2010)

Algoritmo de reconstrução

- Algoritmo de Kruskal adaptado para grafos orientados

M	1	2	3	4	5	6
1	-	5	6	12	46	32
2	26	-	23	23	31	34
3	22	29	-	32	25	39
4	8	35	12	-	16	13
5	31	7	44	19	-	27
6	19	25	31	44	10	-



Trabahos correlatos – Dias et al. (2011)[4]

- Solução inicial para vídeos
  - *Frames* alinhados conforme a cena
  - Cálculo da matriz de dissimilaridade por *frame*
  - Reconstrução de uma árvore filogenética por *frame*
  - As arestas de maior frequência nas árvores geradas formam a árvore final

[4] Z. Dias, A. Rocha e S. Goldenstein. "Video Phylogeny: Recovering Near-Duplicate Video Relationships". Em: *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*. 2011, pp. 1–6.

Trabalhos correlatos – Dias et al. (2013)[6]

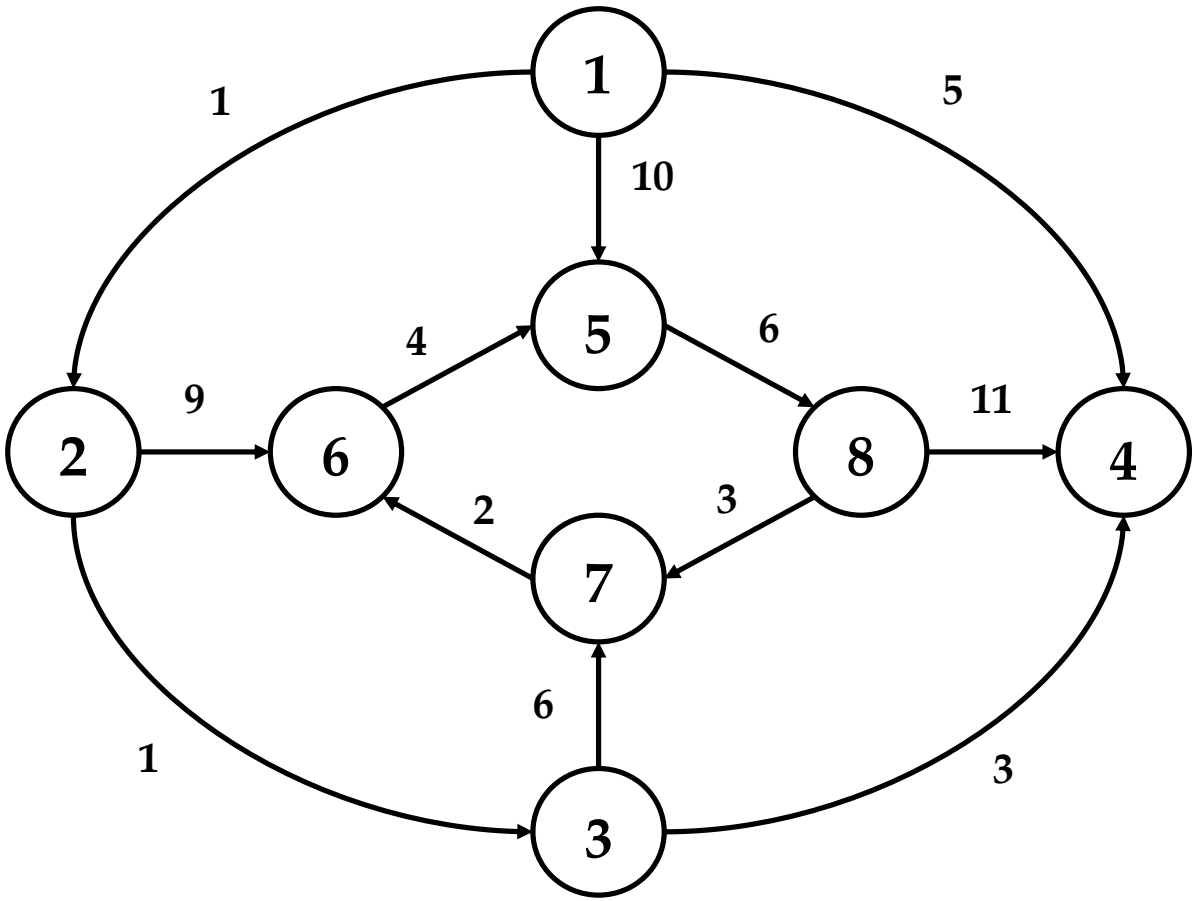
Reconstrução das árvores com o algoritmo para encontrar arborescência ótima em um grafo

- Proposto independentemente por Chu e Liu, Bock e Edmonds[5].

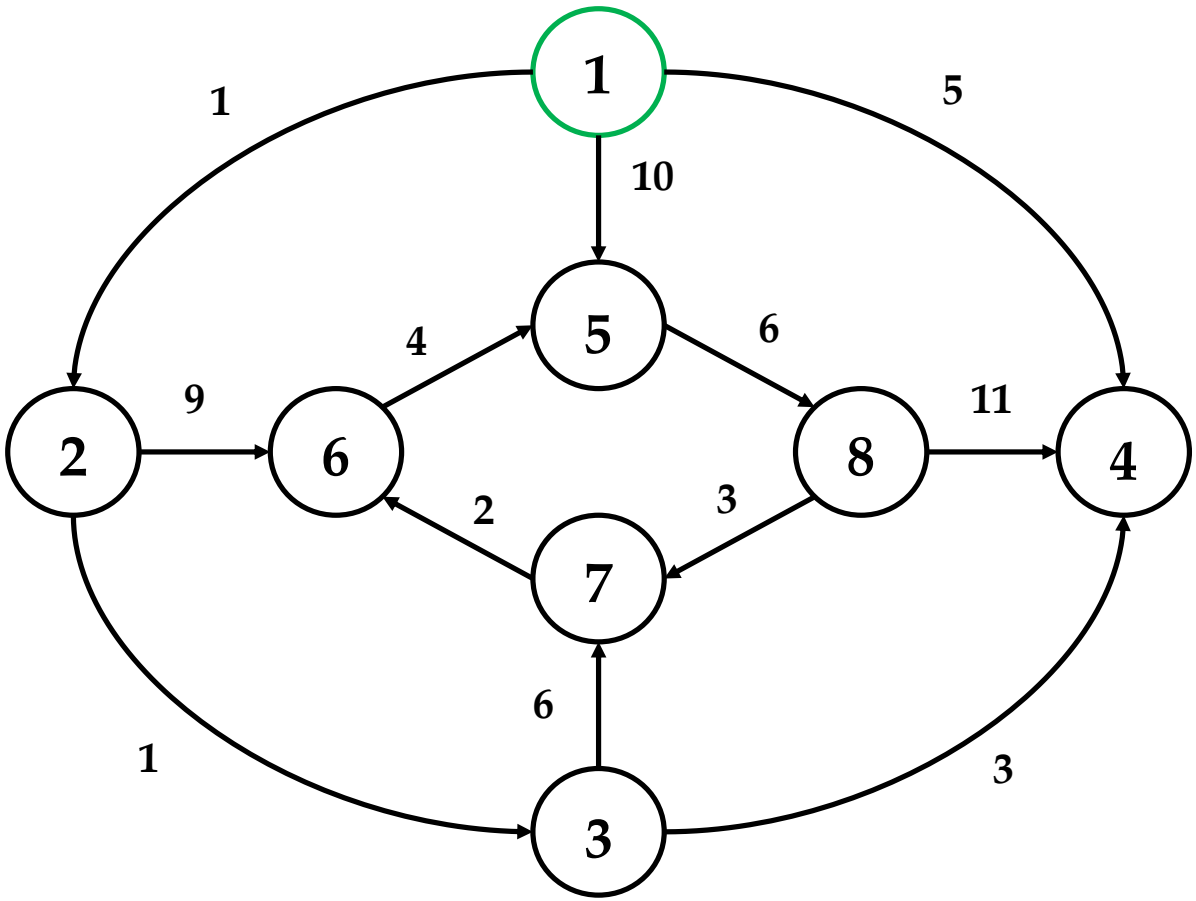
[5] R. E. Tarjan. "Finding optimum branchings". Em: *Networks* 7.1 (1977), pp. 25–35.

[6] Z. Dias, S. Goldenstein e A. Rocha. "Exploring heuristic and optimum branching algorithms for image phylogeny". Em: *Elsevier Journal of Visual Coimmunication and Image Representation* 24.7 (2013), pp. 1124–1134.

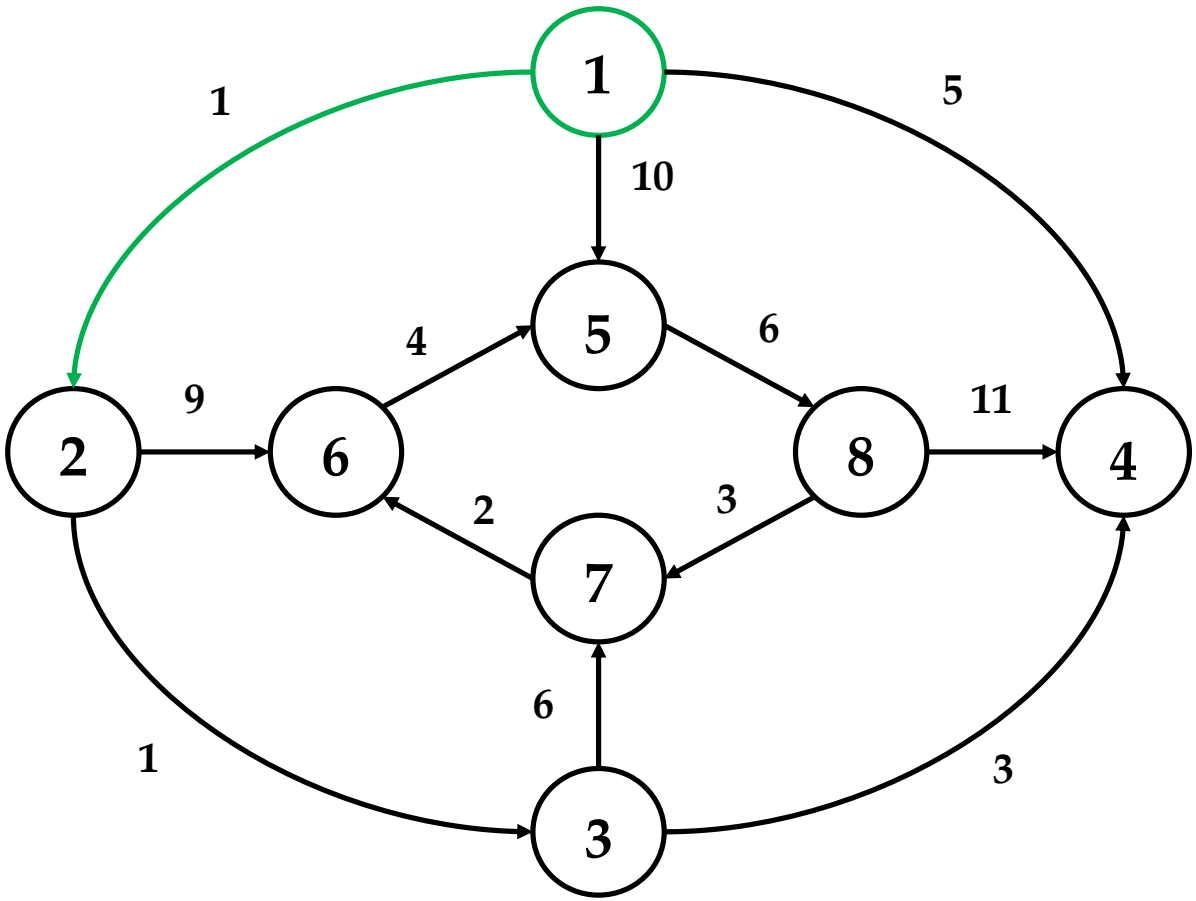
Algoritmo de arborescência ótima



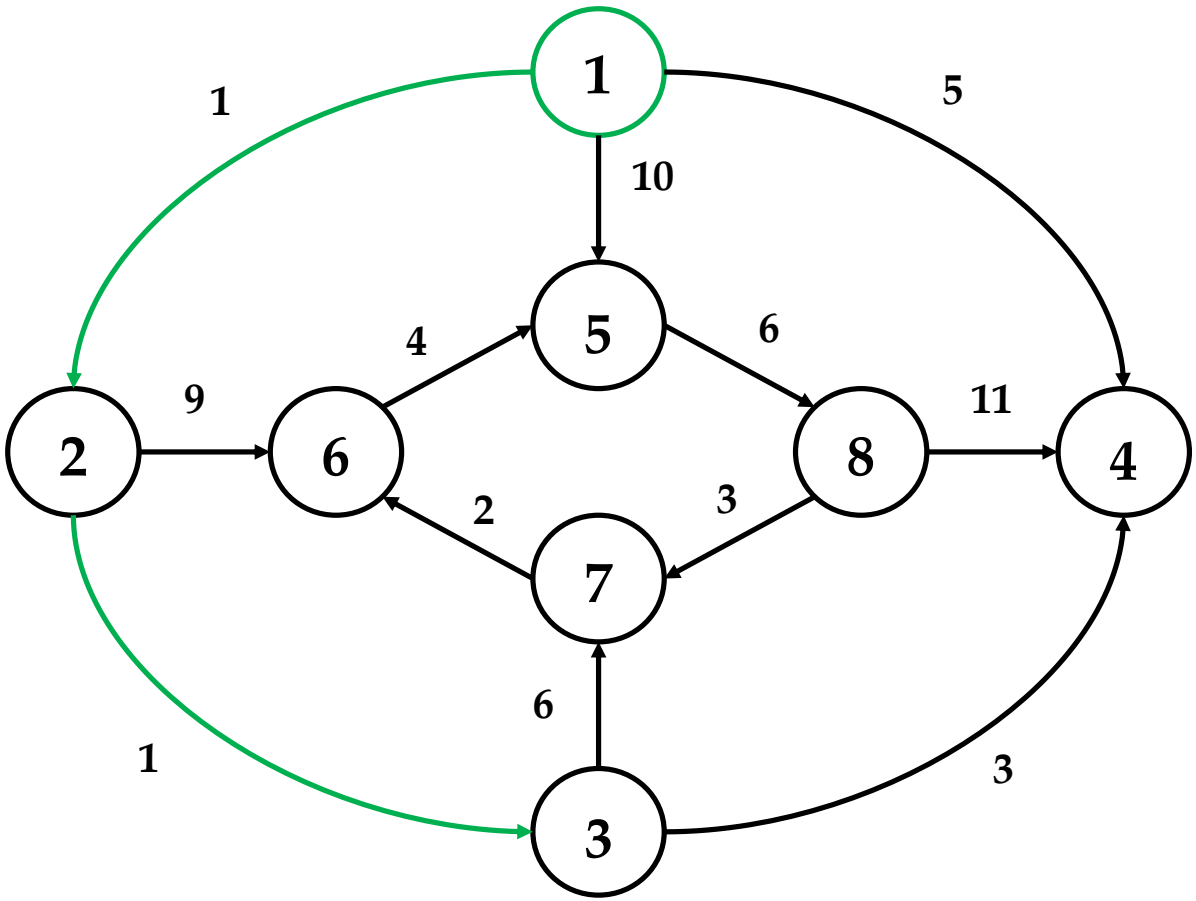
Algoritmo de arborescência ótima



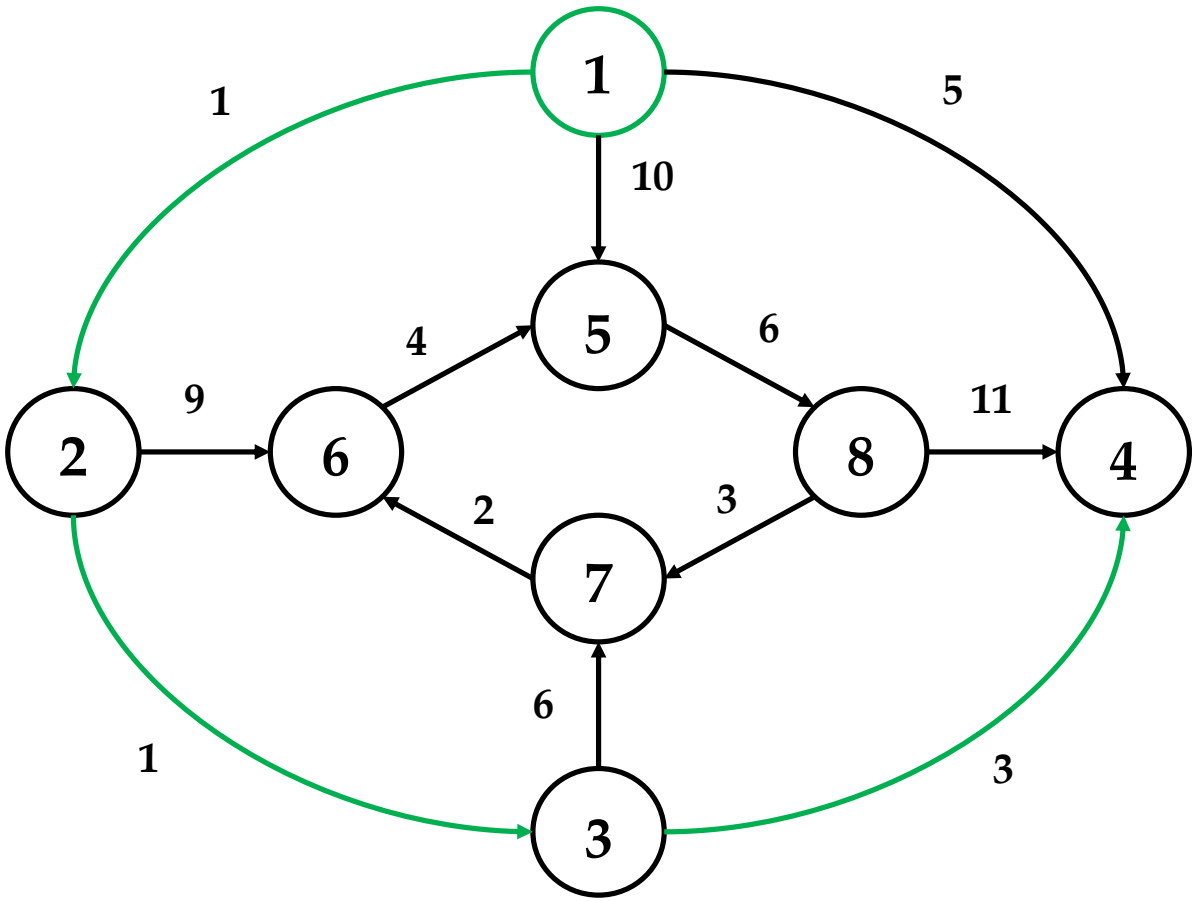
Algoritmo de arborescência ótima



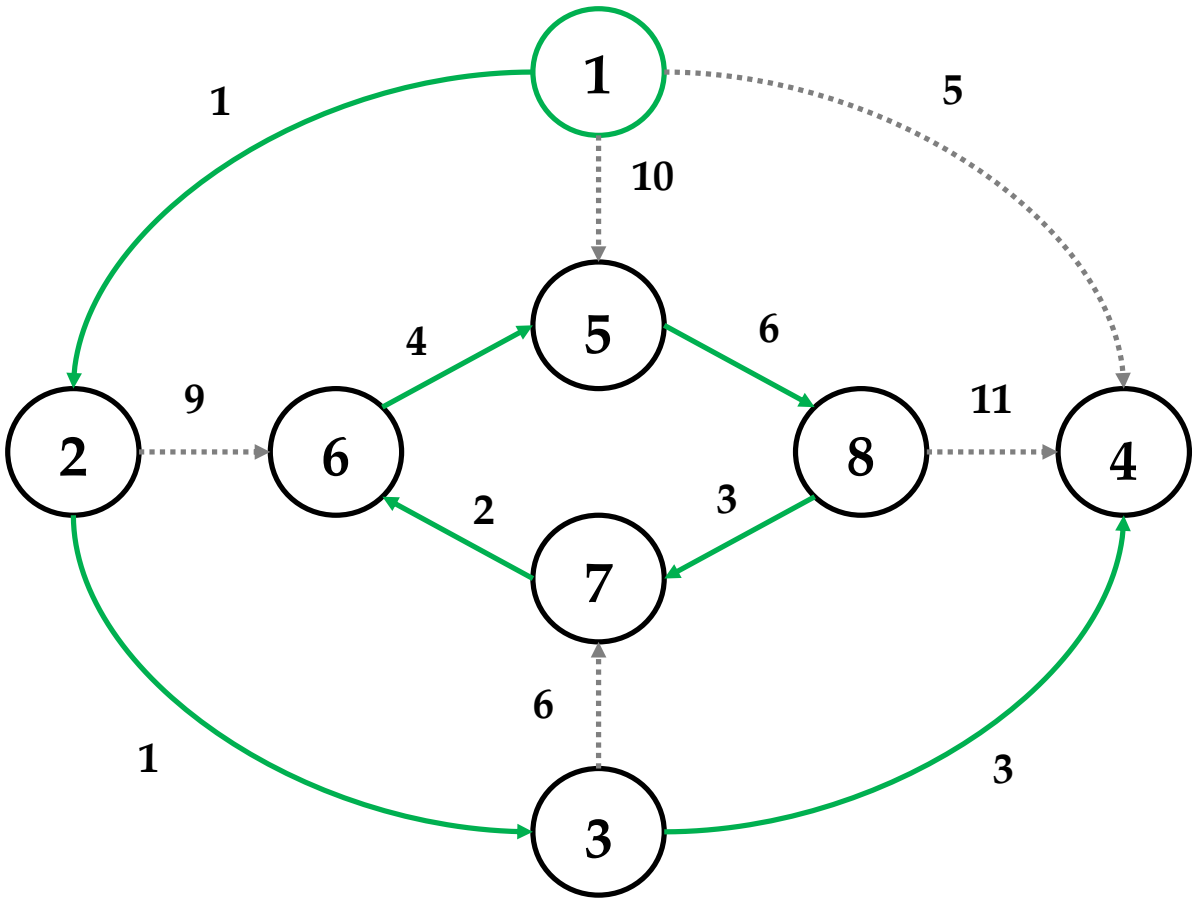
Algoritmo de arborescência ótima



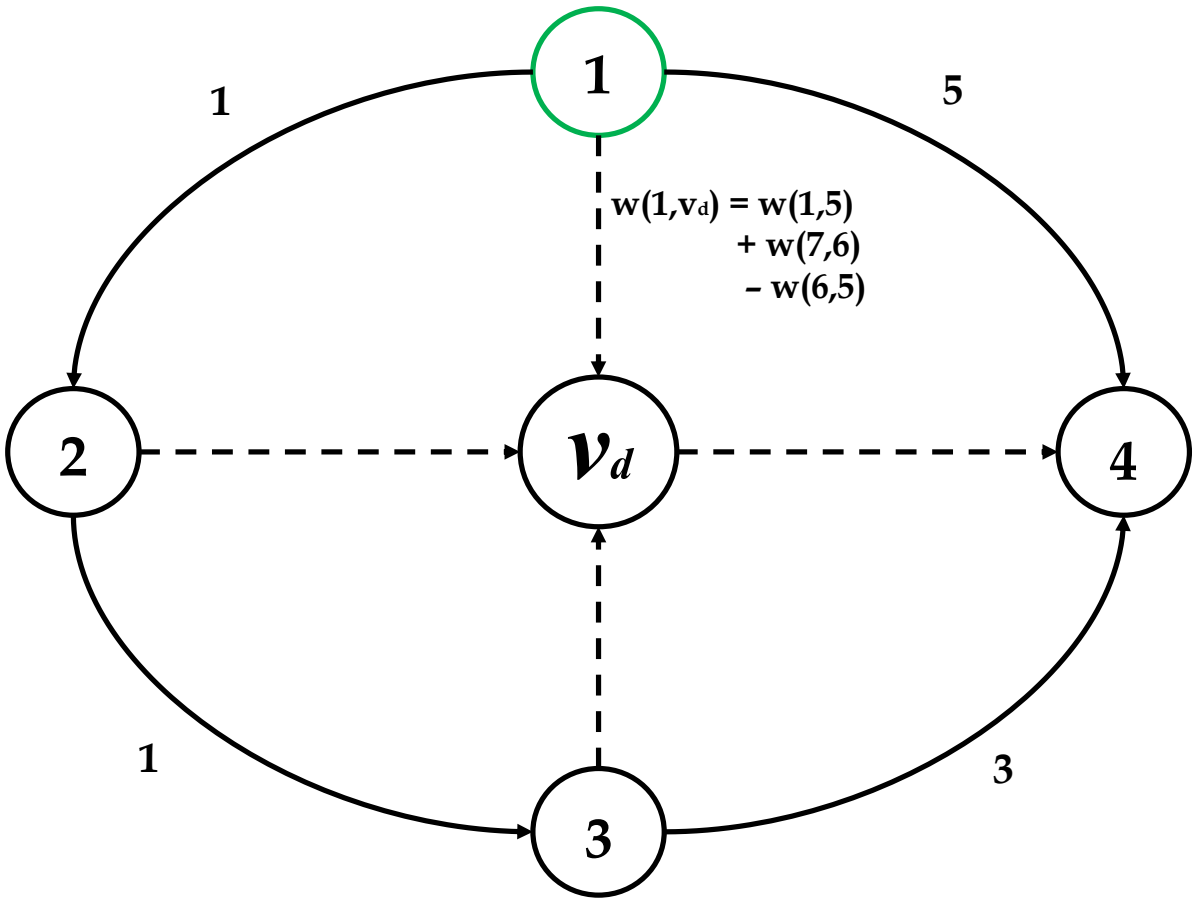
Algoritmo de arborescência ótima



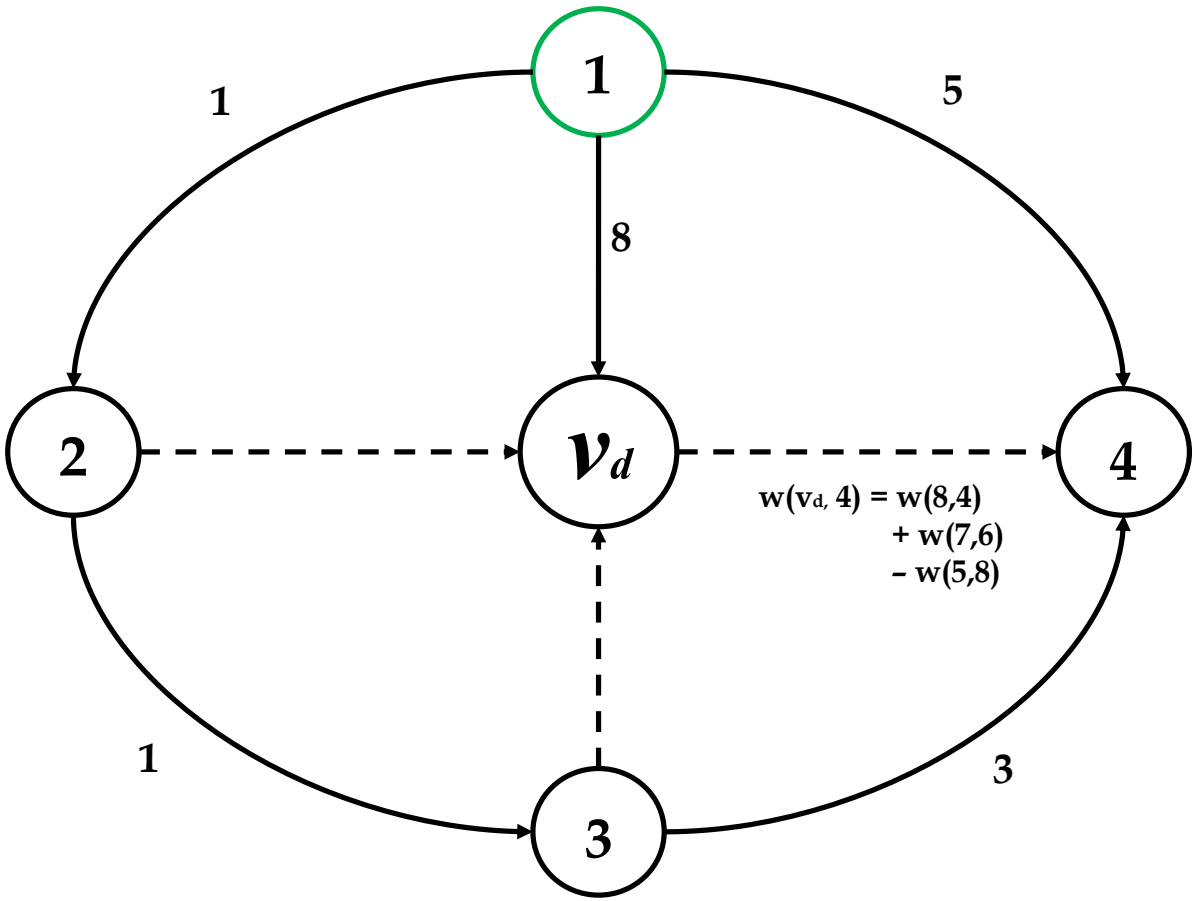
Algoritmo de arborescência ótima



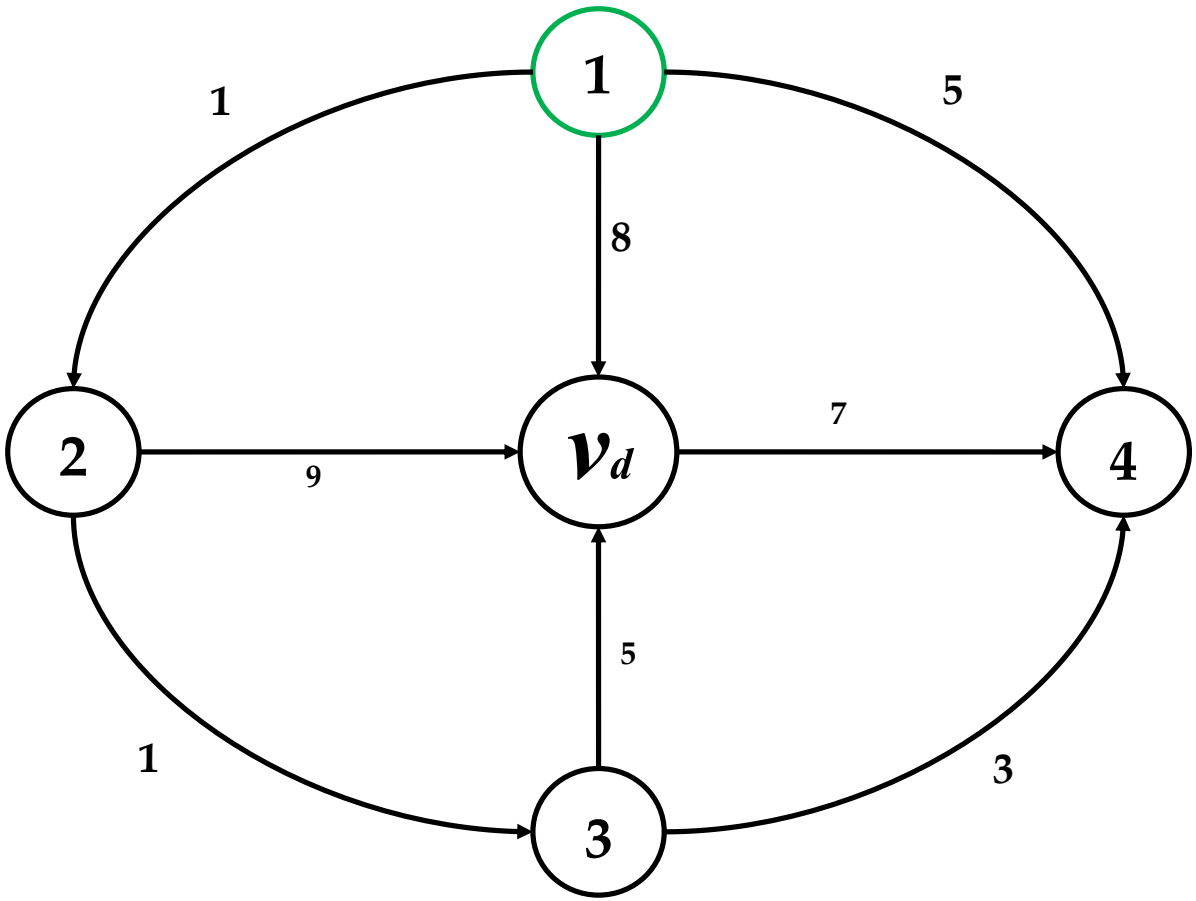
Algoritmo de arborescência ótima



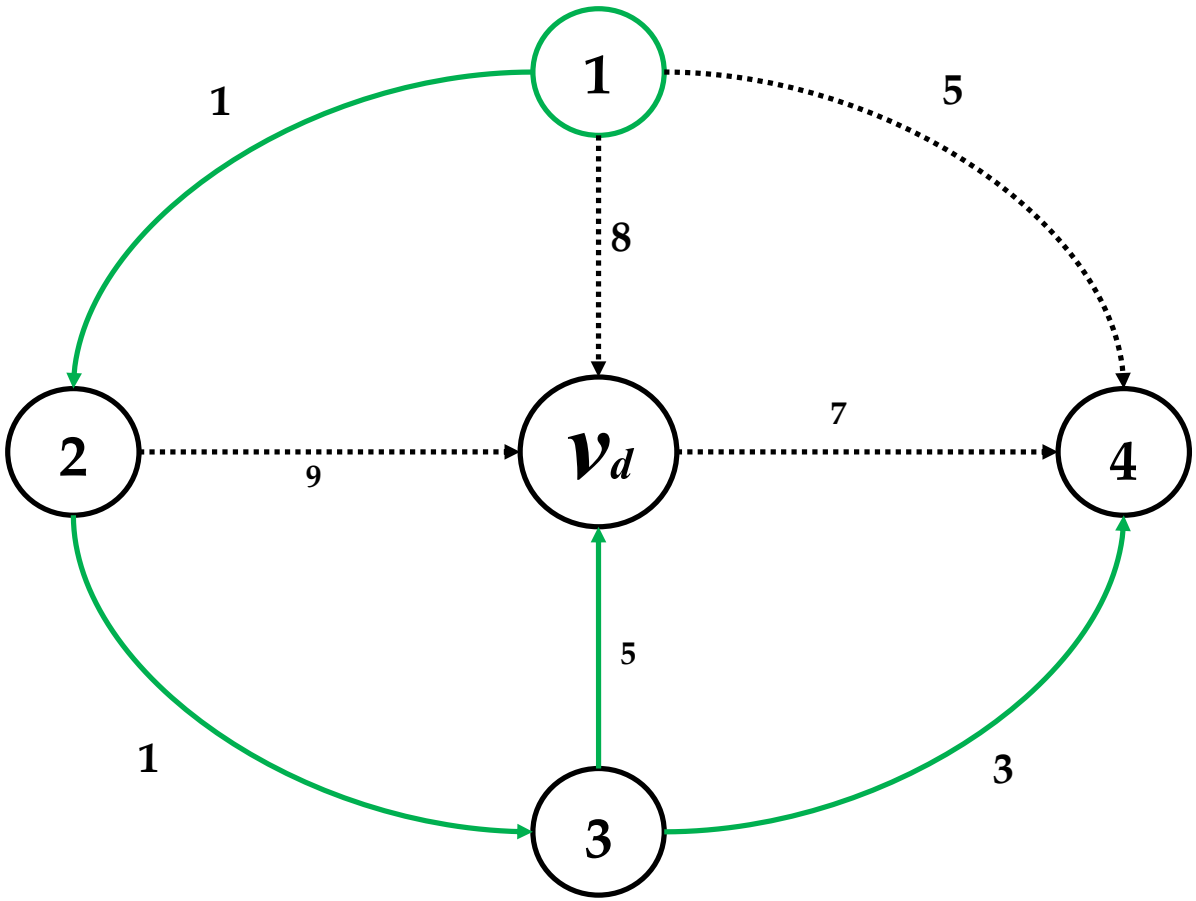
Algoritmo de arborescência ótima



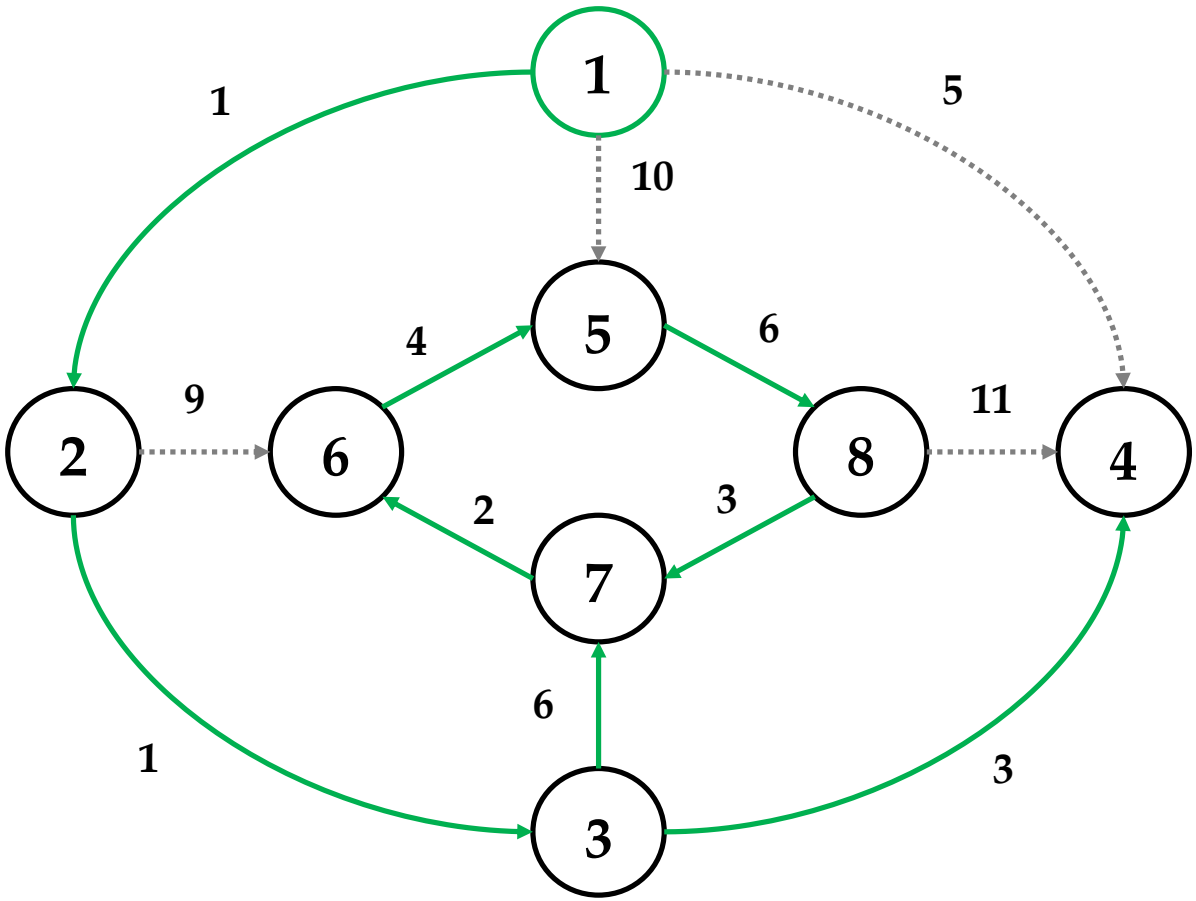
Algoritmo de arborescência ótima



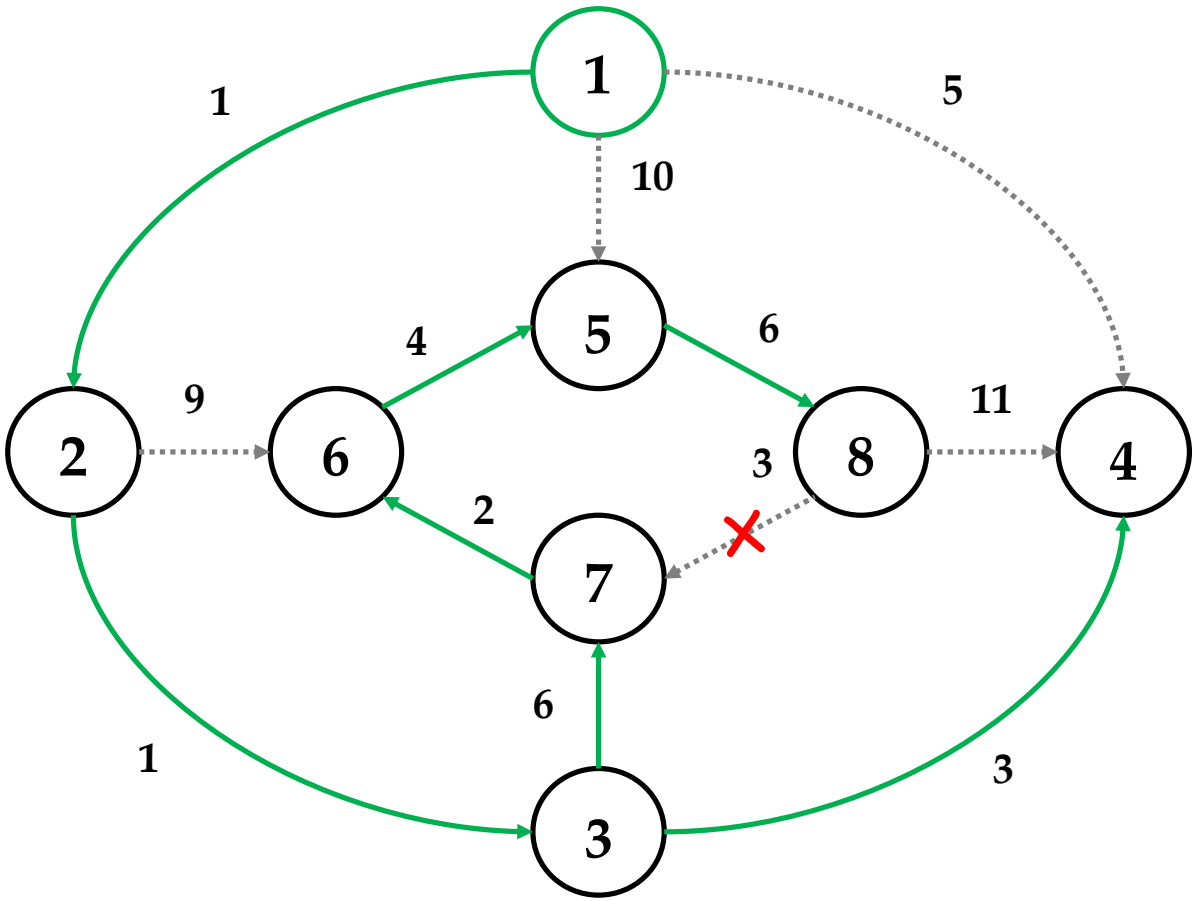
Algoritmo de arborescência ótima



Algoritmo de arborescência ótima



Algoritmo de arborescência ótima



Trabalhos correlatos – Nucci et al. (2013)[7]

- Solução inicial para filogenia de áudio
- Transformações: Codificação, corte, *fade in* e *fade out*
- Estimativa dos parâmetros de transformação: busca em grade
- Dissimilaridade: *Signal Noise Ratio*
- Reconstrução: Kruskal orientado

[7] M. Nucci, M. Tagliasacchi e S. Tubaro. "A phylogenetic analysis of near duplicate audio tracks". Em: *IEEE Intl. Workshop on Multimedia Signal Processing (MMSP)*. 2013.

Trabalhos correlatos – Dias et al. (2013)[8]

- Solução para florestas
  - *Automatic Oriented Kruskal* - AOK
- Para cada aresta a ser adicionada, calcula-se $\tau = c \times \sigma$
  - $\sigma =$ Desvio padrão do peso das arestas já adicionadas
  - $c =$ Constante determinada empiricamente
- Calcula diferença δ entre a aresta a ser adicionada e a última aresta adicionada
- Se $\delta > \tau$, o algoritmo retorna a floresta encontrada até o momento

[8] Z. Dias, S. Goldenstein e A. Rocha. "Toward image phylogeny forests: Automatically recovering sematically similar image relationships". Em: *Forensic Science International* 231 (2013), pp. 178–189.

Metodologia e Desafios

Metodologia

Conjunto de imagens A:

- Duplicatas geradas por uma ou múltiplas câmeras
- 3 cenas
- 3 imagens por câmera
- Quatro tamanhos de florestas ($|F| \in \{2..5\}$)
- Conjuntos de treino e teste
  - Treino: 1 topologia, 10 variações, 2160 florestas
  - Teste: 4 topologias, 10 variações/topologia, 8640 florestas

Metodologia

Conjunto de imagens B:

- Duplicatas geradas por uma ou múltiplas câmeras
- 20 cenas
- 10 imagens por câmera
- 9 tamanhos de florestas ($|F| \in \{2..10\}$)
- 10 topologias para florestas
- 10 variações para cada topologia

Será utilizado somente para a etapa de testes (*Cross Dataset*)

Metodologia

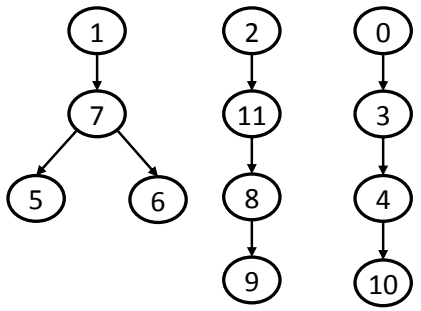
- Cálculo de dissimilaridade
  - Para imagens e vídeos: Dias et al. (2012)
  - Para áudio: Nucci et al. (2013)
- Desafio: existe alguma forma mais eficaz de calcular a dissimilaridade entre pares de duplicatas?

Metodologia

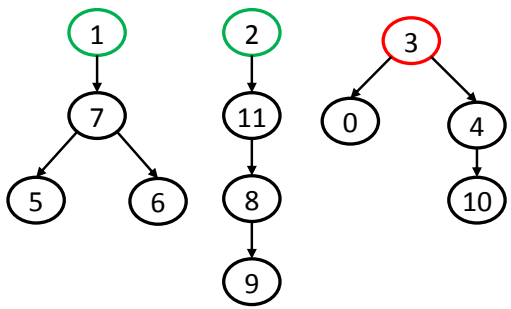
- Métricas de avaliação
  - Raízes
  - Arestas
  - Folhas
  - Ancestrais

Metodologia

Métricas de avaliação – Raízes



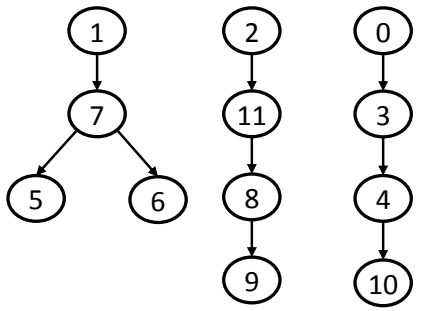
Ground Truth



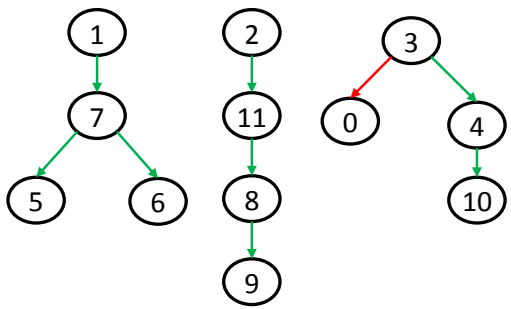
Floresta encontrada

Metodologia

Métricas de avaliação – Arestas



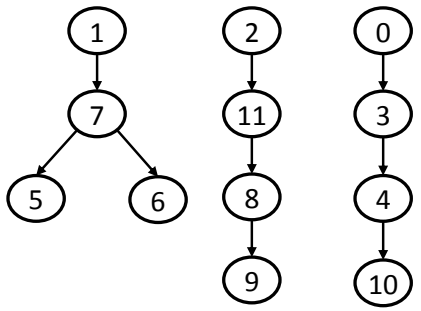
Ground Truth



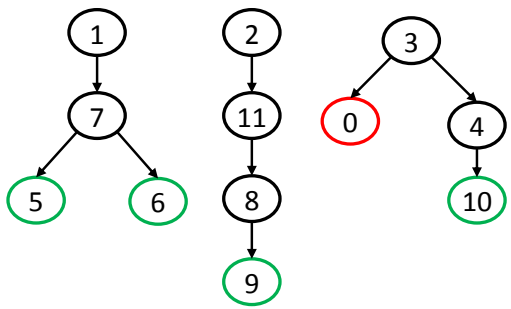
Floresta encontrada

Metodologia

Métricas de avaliação – Folhas



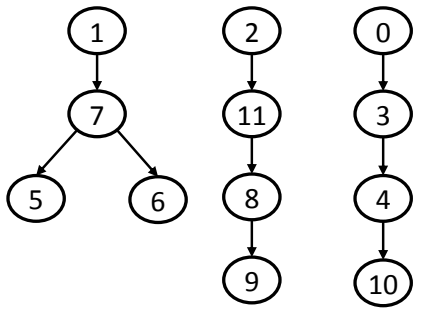
Ground Truth



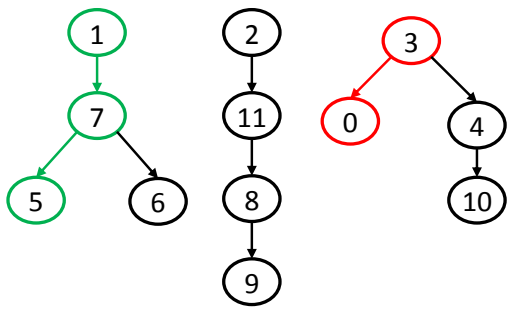
Floresta encontrada

Metodologia

Métricas de avaliação – Ancestrais



Ground Truth



Floresta encontrada

Metodologia

Algoritmos de reconstrução

- AOK
- Arborescência mínima

Desafios:

- Descobrir o número de árvores da floresta
- Descobrir quais documentos farão parte de cada árvore
- Identificar os documentos originais (raízes) da floresta
- Definir a estrutura da floresta

Metodologia

- Experimentos em ambientes controlados e não controlados
- Desafios:
  - Quantos dos documentos são originais?
  - Quais são originais?
  - Qual a relação de dependência entre as duplicatas?
  - Como avaliar uma floresta em um ambiente não controlado?

O que já foi feito?

Reconstrução de florestas filogenéticas

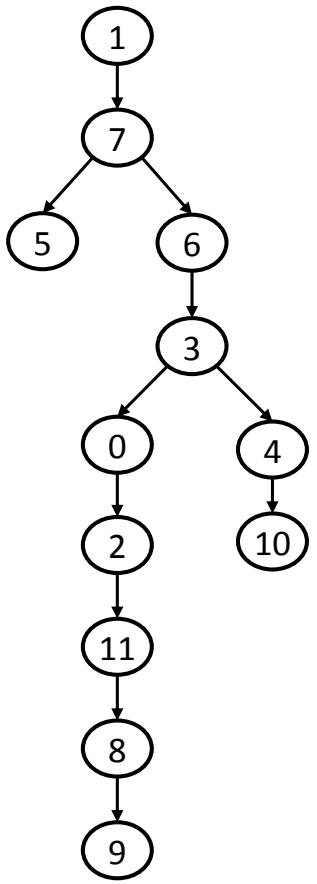
- *Automatic Optimum Branching* – AOB
- AOB estendido
- Fusão de algoritmos de reconstrução

Automatic Optimum Branching (AOB)

Matriz de dissimilaridade

M	0	1	2	3	4	5	6	7	8	9	10	11
0	-	35,12	6,27	1,81	2,98	45,39	36,46	36,19	35,31	36,18	3,85	17,47
1	31,81	-	25,42	28,35	28,36	3,50	3,27	2,88	39,99	40,39	29,09	25,47
2	43,87	48,56	-	39,01	39,06	59,13	49,14	48,18	18,52	19,11	40,63	1,86
3	2,71	34,61	6,39	-	1,88	44,65	35,70	34,22	35,47	36,06	2,74	17,34
4	3,98	34,68	6,86	2,56	-	44,87	35,53	35,19	35,39	36,35	2,50	17,35
5	31,48	4,04	25,72	28,19	28,62	-	2,50	2,69	39,29	40,26	29,06	25,03
6	31,49	3,35	25,92	28,14	28,74	2,43	-	2,37	39,83	40,42	29,13	25,21
7	31,61	3,94	26,35	31,66	28,75	1,45	1,68	-	39,86	40,41	29,14	24,83
8	45,88	49,58	25,57	40,59	40,71	60,87	50,42	49,72	-	2,81	42,71	3,35
9	46,08	49,99	25,77	40,40	40,74	60,52	50,69	50,14	3,12	-	42,71	4,56
10	22,43	35,65	8,08	19,62	19,44	45,79	36,37	36,12	34,99	35,63	-	17,11
11	46,55	50,27	26,35	41,33	41,25	61,15	50,93	50,54	2,96	7,27	43,30	-

Arborescência mínima



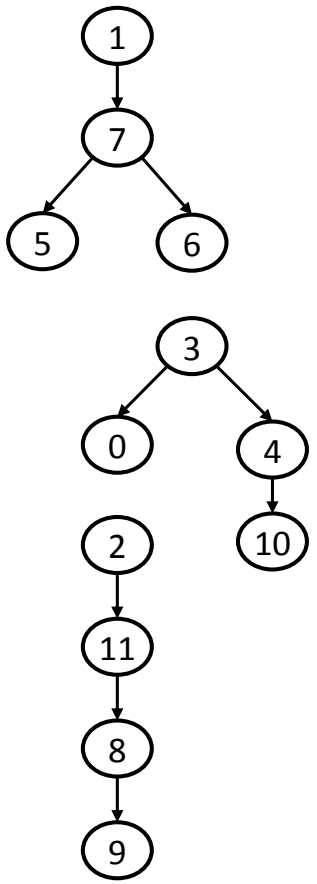
Automatic Optimum Branching (AOB)

Arestas ordenadas

Passo	i →	j	M (i, j)	sd	Limiar (2 x σ)
1	7	5	1,45	–	–
2	7	6	1,68	–	–
3	2	11	1,86	–	–
4	3	4	1,88	–	–
5	4	10	2,5	–	–
6	3	0	2,71	–	–
7	8	9	2,81	0,48	0,10 < 0,97
8	1	7	2,88	0,53	0,07 < 1,07
9	11	8	2,96	0,56	0,08 < 1,12
10	0	2	6,27	0,58	3,31 > 1,16
11	6	3	28,13	–	–

Floresta resultante: [3, 1, 2, 3, 3, 7, 7, 1, 11, 8, 4, 2]
Custo: 20,73

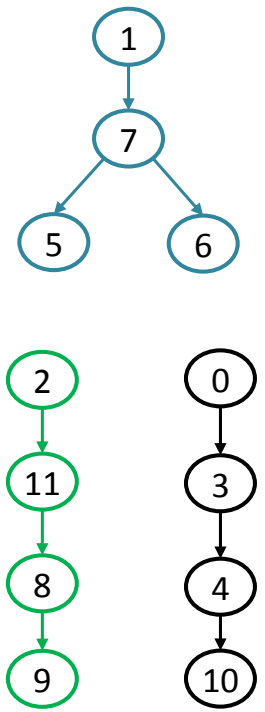
Floresta



AOB Estendido

Matriz de dissimilaridade atualizada

M'	0	3	4	10	1	5	6	7	2	8	9	11
0	-	1,81	2,98	3,85	-	-	-	-	-	-	-	-
3	2,71	-	1,88	2,74	-	-	-	-	-	-	-	-
4	3,98	2,56	-	2,5	-	-	-	-	-	-	-	-
10	22,43	19,6	19,44	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	3,5	3,27	2,88	-	-	-	-
5	-	-	-	-	4,04	-	2,5	2,69	-	-	-	-
6	-	-	-	-	3,35	2,43	-	2,37	-	-	-	-
7	-	-	-	-	3,94	1,45	1,68	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	18,52	19,1	1,86
8	-	-	-	-	-	-	-	-	25,57	-	2,81	3,35
9	-	-	-	-	-	-	-	-	25,77	3,12	-	4,56
11	-	-	-	-	-	-	-	-	26,35	2,96	7,27	-



Floresta resultante: [0, 1, 2, 0, 3, 7, 7, 1, 11, 8, 4, 2]
Custo: 19,83

Definição do limiar τ

- $\tau = c \times \sigma_M$
- Parâmetro c
  - Calculado de forma a maximizar o acerto no conjunto de treino
  - Valor calculado utilizado nos testes

Tabela comparativa de algoritmos para florestas

Experimentos utilizando o conjunto de imagens A

Duplicatas geradas por uma câmera

	AOK				**AOB**				**AOB Estendido**			
$\|F\|$	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**
2	0.834	0.788	0.818	0.740	0.842	0.805	0.834	0.755	0.911	0.809	0.839	0.773
3	0.882	0.823	0.827	0.798	0.898	0.833	0.842	0.814	0.938	0.838	0.847	0.829
4	0.870	0.831	0.820	0.826	0.861	0.835	0.827	0.828	0.914	0.839	0.833	0.843
5	0.883	0.780	0.812	0.762	0.887	0.788	0.824	0.778	0.930	0.791	0.829	0.787

Duplicatas geradas por múltiplas câmeras

	AOK				**AOB**				**AOB Estendido**			
$\|F\|$	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**
2	0.830	0.787	0.817	0.739	0.837	0.805	0.833	0.755	0.908	0.809	0.839	0.773
3	0.883	0.822	0.822	0.801	0.873	0.831	0.837	0.811	0.936	0.837	0.845	0.832
4	0.887	0.830	0.817	0.833	0.835	0.830	0.822	0.821	0.925	0.838	0.832	0.846
5	0.898	0.782	0.814	0.775	0.868	0.786	0.824	0.777	0.937	0.791	0.831	0.794

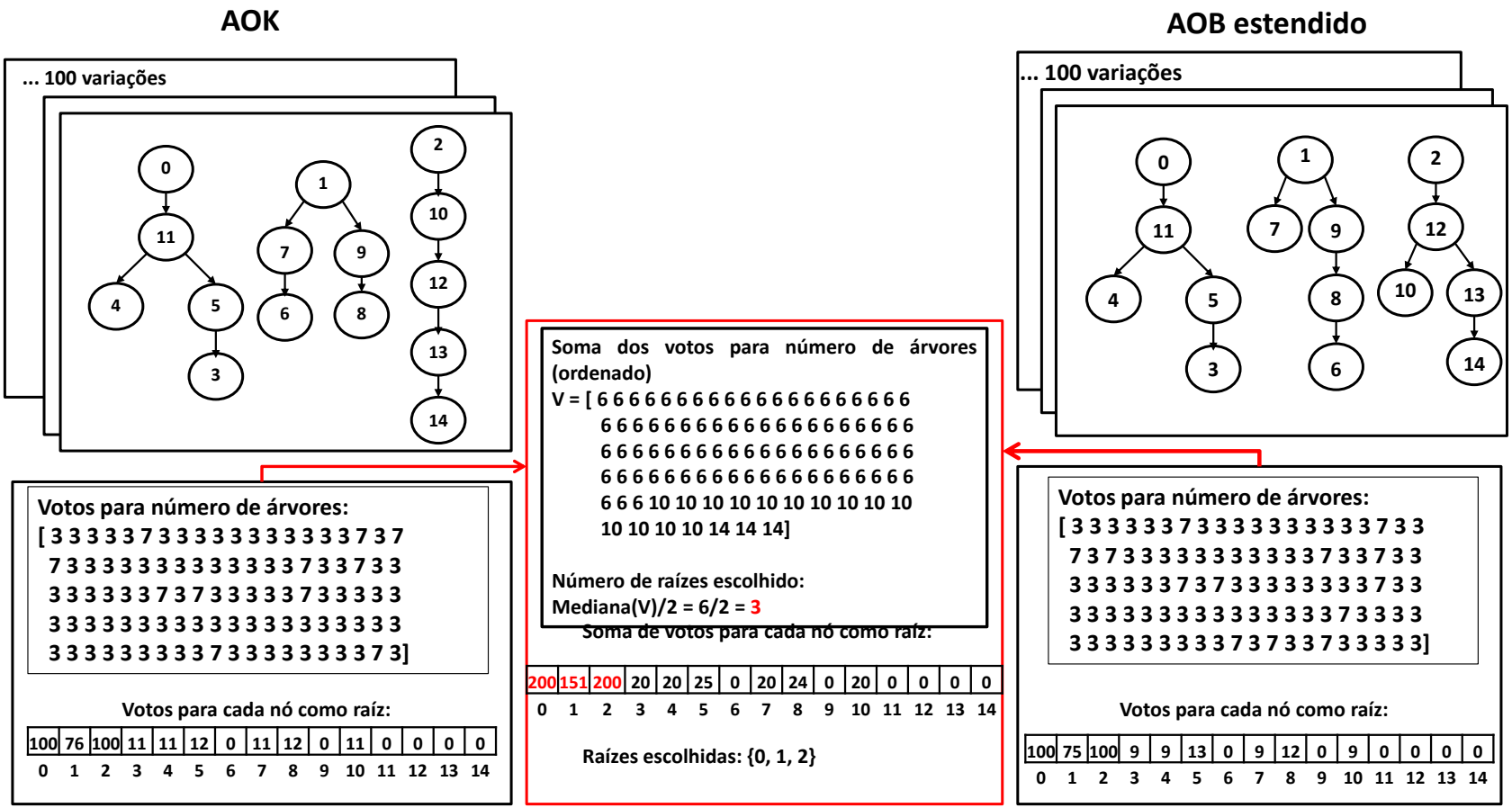
Fusão de algoritmos de reconstrução

- Os algoritmos de reconstrução existentes para floresta apresentam bons resultados
- Existe alguma forma de combinar os algoritmos para melhorar a qualidade da floresta a ser reconstruída?

Fusão de algoritmos de reconstrução

- Perturbação na matriz de dissimilaridade
  - Calcula-se o desvio padrão σ_M de todos os valores da matriz M
  - Para cada posição (i, j) de M, um valor entre $[-k \times \sigma_M, k \times \sigma_M]$ é sorteado e adicionado à $M(i, j)$
    - k é calculado no conjunto de treinamento
- São geradas 100 matrizes

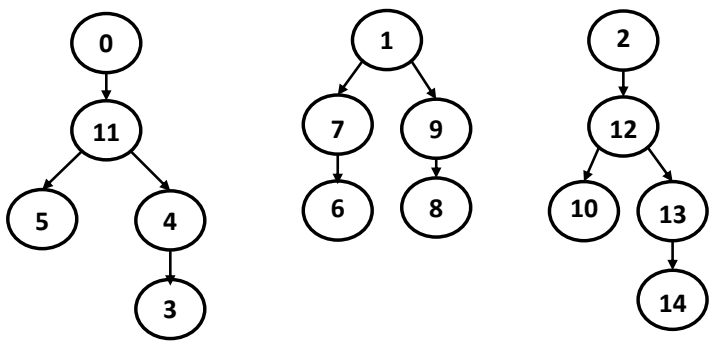
Fusão de algoritmos de reconstrução



Fusão de algoritmos de reconstrução

Soma dos votos das arestas

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	–	–	–	0	0	0	0	0	0	0	0	200	0	0	0
1	–	–	–	0	0	0	0	100	0	199	0	0	0	0	0
2	–	–	–	0	0	0	0	0	0	0	70	0	130	0	0
3	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
4	–	–	–	200	0	0	0	0	0	0	0	0	0	0	0
5	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
6	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
7	–	–	–	0	0	0	200	0	0	0	0	0	0	0	0
8	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0
9	–	–	–	0	0	0	0	0	199	0	0	0	0	0	0
10	–	–	–	0	0	0	0	0	0	0	0	0	70	46	0
11	–	–	–	0	200	199	0	0	0	0	0	0	0	0	0
12	–	–	–	0	0	0	0	0	0	0	130	0	0	154	0
13	–	–	–	0	0	0	0	0	0	0	0	0	0	0	200
14	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0



Floresta final: [0, 1, 2, 4, 11, 7, 1, 9, 1, 12, 0, 2, 12, 13]

Resultados - Fusão

Duplicatas geradas por uma câmera

	AOK × AOB estendido				Redução de erro (AOK)				Redução de erro (AOB estendido)			
$\|F\|$	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**
2	0,924	0,812	0,842	0,780	-54,22%	-11,32%	-13,19%	-15,38%	-14,61%	-1,57%	-1,86%	-3,08%
3	0,940	0,842	0,852	0,833	-49,15%	-10,73%	-14,45%	-17,33%	-3,23%	-2,47%	-3,27%	-2,34%
4	0,923	0,846	0,840	0,849	-40,77%	-8,88%	-11,11%	-13,22%	-10,47%	-4,35%	-4,19%	-3,82%
5	0,931	0,796	0,834	0,790	-41,03%	-7,27%	-11,70%	-11,76%	-1,43%	-2,39%	-2,92%	-1,41%

Duplicatas geradas por múltiplas câmeras

	AOK × AOB estendido				Redução de erro (AOK)				Redução de erro (AOB estendido)			
$\|F\|$	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**	**Raízes**	**Arestas**	**Folhas**	**Ancestrais**
2	0,923	0,811	0,841	0,779	-54,71%	-11,27%	-13,11%	-15,33%	-16,30%	-1,05%	-1,24%	-2,64%
3	0,947	0,843	0,851	0,839	-54,70%	-11,80%	-16,29%	-19,10%	-17,19%	-3,68%	-3,87%	-4,17%
4	0,937	0,844	0,837	0,855	-44,25%	-8,24%	-10,93%	-13,17%	-16,00%	-3,70%	-2,98%	-5,84%
5	0,948	0,796	0,835	0,800	-49,02%	-6,42%	-11,29%	-11,11%	-17,46%	-2,39%	-2,37%	-2,91%

Próximos passos

- Experimentos com o conjunto de imagens B
- Submissão de artigo para uma revista científica
- Experimentos iniciais com filogenia de vídeo

Etapas do projeto

Etapas do projeto

1 - Obtenção de créditos em disciplinas

2 - Revisão da literatura

3 - Geração de conjuntos de duplicatas de imagens

4 - Filogenia de imagens

Etapas do projeto

5 - Qualificação de Doutorado

6 - Programa de Estágio Docente (PED)

7 - Geração de conjuntos de duplicatas de vídeos

8 - Filogenia de vídeos

9 - Doutorado sanduíche no *Politecnico di Milano*, em Milão, Itália

Etapas do projeto

10 - Possíveis abordagens para filogenia de arquivos de áudio

11 - Publicação dos resultados obtidos

12 - Escrita da tese

13 - Defesa da tese de doutorado

Cronograma do projeto

Ano		2012			2013						2014						2015						2016		
Bimestre		4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3
Etapas	1	●	●	●	●	●	●																		
	2	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●						
	3				●	●	●																		
	4					●	●	●	●	●	●														
	5									●															
	6										●	●	●												
	7										●	●	●												
	8										●	●	●	●	●	●	●								
	9													●	●	●									
	10																		●	●	●	●	●		
	11									●			●				●				●				●
	12																				●	●	●	●	●
	13																								●

Agradecimentos

- Instituto de Computação - UNICAMP
- CAPES
- REWIND – União Europeia
- Marina Oikawa
- Professores Anderson Rocha, Zanoni Dias e Siome Goldenstein