1. Introducción

Lo que buscamos con este reporte es poder entender características específicas sobre grafos y sacar conclusiones con respecto a ellos, para lo cual usaremos como herramienta de estudio las Cadenas de Markov.

Las características sobre las que trabajaremos son las siguientes:

- Distribución Estacionaria
- Tiempos de Cruce
- Tiempo de Cubrimiento

En la siguiente sección definiremos los conceptos mencionados anteriormente.

También trabajaremos sobre el modelo PageRank, el cual es una marca registrada y patentada por Google que ampara una familia de algoritmos utilizados para asignar de forma numérica la relevancia de las páginas web indexados por un motor de búsqueda. Dicho modelo es utilizado por el motor de búsqueda Google para determinar la importancia o relevancia de una página.

Por ultimo analizaremos dos estrategias para ayudar a que un Spammer, que agregó su sitio a la web, mejore el ranking de su página.

Vamos a utilizar dos grafos particulares para poder analizar en ellos las características previamente mencionadas.

2. Definiciones, Métodos, Estrategias y Grafos

En esta sección definiremos los conceptos nombrados en la introducción, métodos a usar, estrategias propuestas por el Spammer y los grafos que utilizaremos para concluir resultados.

2.1. Definiciones de conceptos

Tiempo de Cruce: El tiempo de cruce para un nodo es la cantidad de pasos que necesita un caminante aleatorio para volver a pasar por dicho nodo.

Tiempo de Cubrimiento: El tiempo de cubrimiento para un grafo es la cantidad de pasos que necesita un caminante aleatorio para visitar cada nodo del grafo al menos una vez, comenzando en un nodo elegido de forma aleatoria.

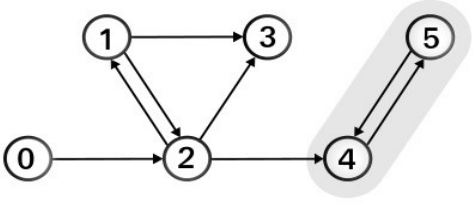
Matriz de Transición: Es una matriz cuadrada de tamaño n (cantidad de nodos del grafo) la cual nos indica los vecinos y la probabilidad de movernos a ellos.

Tendremos dos matrices de transición para cada grafo. Las nombraremos de la siguiente forma:

- **Matriz de Transición Original:** Es una matriz de transición, en la cual la probabilidad de pasar a un vecino es equiprobable, es decir, que la probabilidad es la misma para todos los vecinos, como por ejemplo la Matriz de Transición P de la Figura 1.

- **Matriz de Transición Modificada:** Es una matriz de transición generada por el modelo PageRank que tiene un factor de dumping α (alfa) entre 0 y 1. La intuición de dicho modelo es la de capturar el comportamiento de un "navegador aleatorio" que con probabilidad (1-α) se aburre y efectúa un salto a un sitio arbitrario.

  Este modelo genera una matriz de transición que soluciona el problema que tendría la Matriz de Transición Original, si hubiera "nodos colgados" (por ejemplo, nodo 3 en la Figura 1), nodos que no puedan ser alcanzados en un número finito de pasos (por ejemplo, nodo 0 en la Figura 1) o la existencia de bucles absorbentes (por ejemplo, nodos 4-5 en la Figura 1).



$$P = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figura 1: Grafo y Matriz de Transición de ejemplo.

2.2. Métodos para la obtención de la Distribución Estacionaria

Para la obtención de la distribución estacionaria se utilizaron dos métodos:

- **Método 1:** Se empleó un caminante aleatorio para que se moviera sobre el grafo según las probabilidades que figuren en la Matriz de Transición. Una vez que se realicen X pasos, el caminante se detendrá, de ahí tomaremos el nodo en el que se detuvo. Realizando esto N veces, es decir, hacer N caminatas sobre el grafo podremos determinar la Distribución Estacionaria de dicho grafo como la cantidad de caminantes que terminaron en un nodo sobre el total de N caminatas hechas.

- **Método 2:** Un forma analítica para determinar la Distribución Estacionaria es usar el "Método de las Potencias", un método iterativo pero con una baja velocidad de convergencia. Este método consiste en tomar una distribución inicial arbitraria y una Matriz de Transición, y a partir de estos datos calcular de forma iterativa durante una serie de X pasos la Distribución Estacionaria.

2.3. Estrategias propuestas por el Spammer

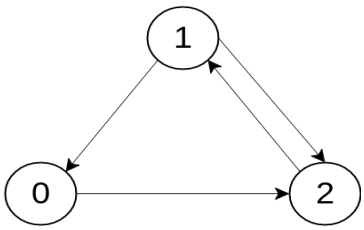
Como se dijo en la introducción, un Spammer desea que su sitio web obtenga el mayor ranking posible, el mismo elabora dos estrategias para mejorar el ranking:

- **Estrategia A:** Crear K páginas ficticias y hacer que todas ellas apunten a su sitio. Después de esta adición, el numero total de páginas de la web (suponiendo que había n páginas) sera de n + K + 1 páginas.

- **Estrategia B:** Hackear K páginas de las n existentes, agregándoles enlaces a su página.

2.4. Grafos a utilizar y su representación

Primero que nada aclararemos cómo serán representados los grafos en este reporte. Por ejemplo, el siguiente grafo :



Se representa de la siguiente manera:

```
G = ( (2,) , (0, 2) , (1,) )
```

Donde cada índice de la tupla representa un nodo y a su vez cada nodo indica los vecinos a los cuales puede desplazarse.

Como se dijo en la introducción, utilizaremos dos grafos particulares para su posterior análisis. A continuación nombraremos dichos grafos:

Grafo 1: Tiene un total de 10 nodos y se representa de la siguiente manera:

```
G1 = ( (2,) , (0, 5, 3) , (6, 0, 2, 7, 4) , (3, 1) ,
       (1, 4, 2, 9, 5, 6, 3, 8, 7) , (0, 4, 5) , (6, 9, 0) ,
       (3, 9) , (7, 2, 6, 5, 3, 1, 0) , (2, 0) )
```

Para referirnos a este grafo usaremos la abreviación: **G1**.

Grafo 2: Tiene un total de 100 nodos. Debido a su gran extensión no detallaremos como se representa. Para referirnos a este grafo usaremos la abreviación: **G2**.

Vale aclarar que ambos grafos nombrados son conexos.

3. Algoritmos

En esta sección mostraremos los pseudocódigos de los algoritmos que fueron empleados para la obtención de los resultados.

NOTA: En algunos de los pseudocódigos se nombra a alguna de las siguientes funciones:

- g2p_pagerank
- power_iter_one_step

Dichas funciones ya son conocidas, están dadas y definidas; por lo que no daremos su pseudocódigo en este reporte.

3.1. Pseudocódigo Método 1

```
function distribucionEstacionaria(grafo, matriz_trans, pasos, simulaciones):
    length_grafo ← len(grafo) # Tamaño del grafo
    apariciones ←Generar vector de tamaño length_grafo con todos 0's
    dis_est ←Generar vector de tamaño length_grafo con todos 0's

    for 1 to simulaciones:
        nodo ← Elegir un nodo aleatorio del grafo
        for 1 to pasos-1:
            vecino ← Elegir un vecino del nodo elegido de forma aleatoria
            nodo ← vecino
        end for
        apariciones[nodo] ← apariciones[nodo] + 1
    end for
    for i in length_grafo:
        dist_est[i] ← apariciones[i] / simulaciones
    end for

    return dist_est

end function
```

3.2. Pseudocódigo Método 2

```
function metodoPotencias(grafo, matriz_trans, pasos):
    length_grafo ← len(grafo) # Tamaño del grafo
    pi ← Generar vector de tamaño length_grafo con todos 1's
    for 1 to pasos:
       pi ← power_iter_one_step(pi, matriz_trans) # Realiza un paso del método de las potencias
    end for

    return pi

end function
```

3.3. Pseudocódigo Tiempo de cruce

```
function visitarNodo(nodo, matriz_trans, length_grafo, simulaciones)
    t ← 0
    for 1 to simulaciones:
       nodo_inicial ← nodo
       while True:
          vecino ← Elegir un vecino del nodo_inicial elegido de forma aleatoria
          t ← t + 1
          # Si el vecino es el nodo con el cual ya empeze corto la búsqueda
          if nodo = vecino:
             break
          end if
          nodo_inicial ← vecino
        end while
    end for
   pasos ← t / simulaciones # Promedio de pasos para cruzarme de nuevo con el nodo

   return pasos

end function



function tiempoCruce(grafo, matriz_trans, simulaciones):
    length_grafo ← len(grafo) # Tamaño del grafo
    tiempos ← []
    media ← 0
    # Calculamos el tiempo de cruce para todos los nodos del grafo
    for nodo in grafo:
        Agregar (visitarNodo(nodo, matriz_trans, length_grafo, simulaciones)) a tiempos
    end for

    media ← (Sumar todos los elementos de tiempos) / simulaciones # Media de tiempos

    return tiempos, media

end function
```

3.4. Pseudocódigo Tiempo de cubrimiento

```
function tiempoCubrimiento(length_grafo, matriz_trans):
    nodo ← Elegir un nodo aleatorio del grafo
    nodos_vistos ← [nodo] # nodos por el caminante paso
    pasos ← 1

    # Corremos el caminante aleatorio hasta pasar por todos los nodos
    while len(nodos_vistos) < length_grafo:
       nodo ← Elegir un vecino del nodo elegido de forma aleatoria
       pasos ← pasos + 1
       if nodo not in nodos_vistos:
          Agregar  (nodo) a nodos_vistos
       end if
    end while

    return pasos

end function

function cubrirGrafo(grafo, matriz_trans, simulaciones):
    length_grafo ← len(grafo) # Tamaño del grafo
    pasos ← 0
    for 1 to simulaciones:
       pasos ←  pasos + tiempoCubrimiento(length_grafo, matriz_trans)
    end for
    tiempo ← pasos / simulaciones

    return tiempo

end function
```

3.5. Pseudocódigo Estrategia A

```
function paginasFictias(grafo, K):
    n ← len(grafo) # Tamaño del grafo
    Agregamos un nodo vacío al grafo # Mi pagina web
    for 1 to K:
        Agregamos un nodo que apunte al nodo n+1 al grafo # Agrego K paginas a la web
    end for

    N ← n + K + 1 # Nuevo tamaño del grafo
    P ← g2p_pagerank(grafo, 0.85) # Matriz de transición con PageRank con α = 0.85
    dist_est ← metodoPotencias(N, P, 100)
    ranking ← dist_est[n]

    return ranking

end function
```

3.6. Pseudocódigo Estrategia B

```
function hackearPaginas(grafo, K):
    n ← len(grafo) # Tamaño del grafo
    Agregamos un nodo vacío al grafo # Mi pagina web
    nodos_hackeados ← []
    for 1 to K:
        nodo ← nodoAleatorio(n)
        while nodo in nodos_hackeados:
            nodo ← Elegir un nodo aleatorio del grafo
        end while
        Agregar (nodo) a  nodos_hackeados
        Agregamos un enlace al nodo hacia el nodo n+1
   end for

   N ← n 1 # Nuevo tamaño del grafo
   P ← g2p_pagerank(grafo, 0.85) # Matriz de transición con PageRank con α = 0.85

   dist_est ← metodoPotencias(N, P, 100)
   ranking ← dist_est[n]

   return ranking

end function
```

4. Resultados

En esta sección se mostraran los resultados obtenidos sobre:

- La Distribución Estacionaria sobre G1.
- Los Tiempos de Cruce sobre G1 y G2.
- El Tiempo de Cubrimiento sobre G1 y G2.

4.1. Distribución Estacionaria

Para el calculo de la Distribución Estacionaria utilizamos la Matriz de Transición Original y la Modificada ($\alpha = 0.85$) con ambos métodos mencionados anteriormente.

Para el Método 1 se utilizaron 100 pasos para el caminante aleatorio y para el Método 2 se utilizaron 100 pasos del método de las potencias.

Representaremos la Distribución Estacionaria como una lista, donde cada posición de la lista se corresponde con cada nodo del grafo.

4.1.1. Distribución Estacionaria con Método 1

Matriz de Transición Original:

[0.162 , 0.077 , 0.262 , 0.133 , 0.078 , 0.055 , 0.094 , 0.062 , 0.009 , 0.068]

Matriz de Transición Modificada ($\alpha = 0.85$):

[0.155 , 0.081 , 0.233 , 0.127 , 0.085 , 0.067 , 0.088 , 0.067 , 0.025 , 0.072]

4.1.2. Distribución Estacionaria con Método 2

Matriz de Transición Original:

[0.163 , 0.077 , 0.261 , 0.133 , 0.079 , 0.053 , 0.093 , 0.062 , 0.009 , 0.071]

Matriz de Transición Modificada ($\alpha = 0.85$):

[0.157 , 0.081 , 0.229 , 0.132 , 0.081 , 0.068 , 0.09 , 0.064 , 0.023 , 0.075]

Como podemos apreciar a simple vista, las cuatro distribuciones obtenidas son prácticamente iguales, pero hay algunas pequeñas diferencias al usar una u otra Matriz de Transición, dicha diferencia la podemos apreciar mejor viendo la Figura 2.

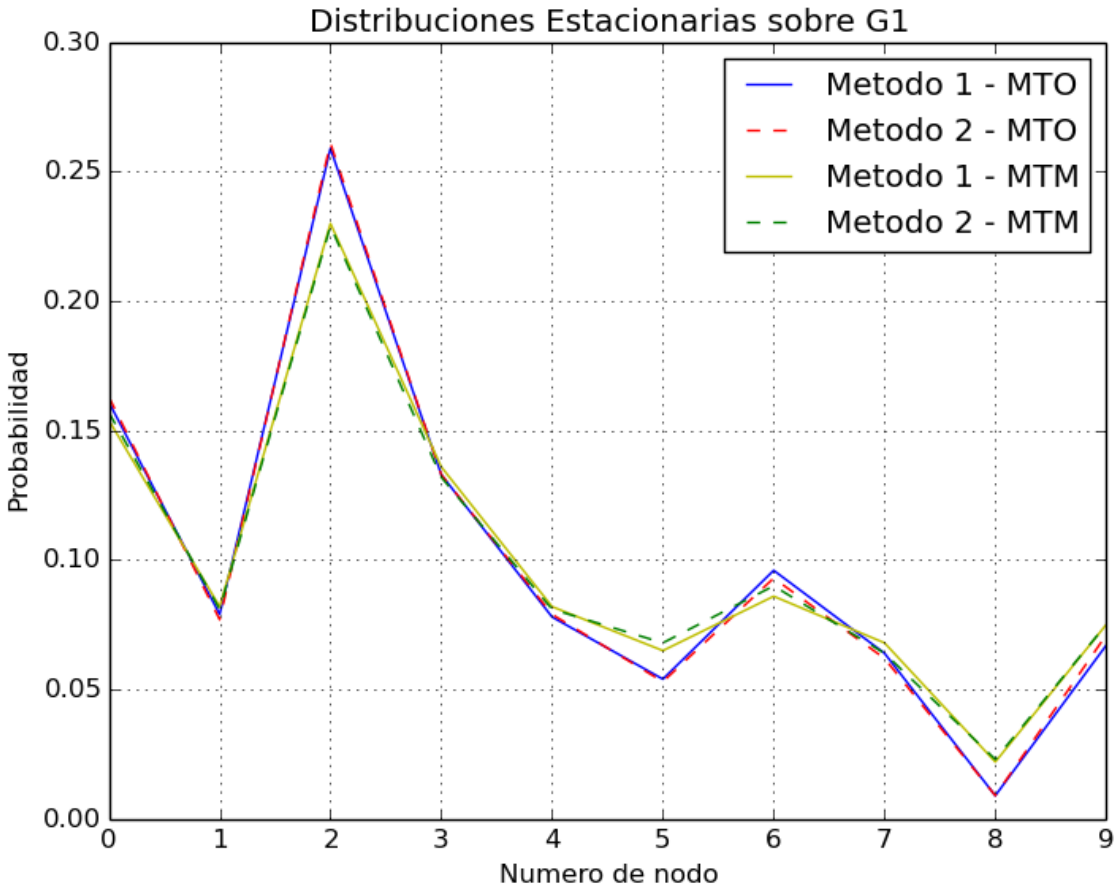


Figura 2: Comparación entre las Distribuciones Estacionarias obtenidas.

Vale aclarar que las abreviaciones MTO y MTM vistas en la Figura 1 y que veremos en algunos gráficos de las secciones siguientes, corresponden a **M**atriz de **T**ransición **O**riginal y **M**atriz de **T**ransición **M**odificada respectivamente.

Podemos notar que las Distribución Estacionaria obtenida con ambos métodos sobre la Matriz de Transición Original son prácticamente iguales. La misma similitud existe al usar ambos métodos sobre la Matriz de Transición Modificada.

Pero a pesar de esto no existen grandes diferencias. Por lo que podemos concluir que ambos métodos son buenos y no debería descartarse ninguno de ellos.

4.2. Tiempos de Cruce

4.2.1. Tiempos de Cruce para G1

A continuación mostraremos los Tiempos de Cruce de cada nodo del grafo G1 y la media de dichos tiempos, usando la Matriz de Transición Original y la Modificada con $\alpha = 0.85$.

Representaremos los Tiempos de Cruce los nodos como una lista, donde cada posición de la lista se corresponde con cada nodo del grafo.

Con Matriz de Transición Original:

[6.14 , 13.39 , 3.48 , 5.07 , 12.83 , 19.2 , 10.71 , 18.02 , 112.73 , 13.27]

Media de Tiempos = 21.953

Podemos notar en la lista de tiempos que el nodo 8 tiene un alto tiempo de cruce, esto quiere decir que durante las caminatas realizadas debe pasar mucho tiempo para volver a pasar sobre dicho nodo. Esto genera que el tiempo medio de cruce aumente considerablemente.

Con Matriz de Transición Modificada ($\alpha = 0.85$):

[7.07 , 12.38 , 4.71 , 6.6 , 10.94 , 13.62 , 8.95 , 15.94 , 46.09 , 12.14]

Media de Tiempos = 14.397

Al usar la Matriz de Transición Modificada vemos que casi todos los tiempos de cruce de cada nodo varia muy poco, a diferencia del nodo 8 donde su tiempo disminuye bastante, esto nos dice que al usar esta matriz de transición dicho nodo tendrá una mayor participación en las caminatas realizadas sobre el grafo, esto causa que el tiempo medio de cruce disminuya con respecto al obtenido con la Matriz de Transición Original.

Todo esto que acabamos de decir lo podemos analizar mejor viendo la Figura 3, donde comparamos los tiempos de cruce obtenidos con cada una de la Matrices.

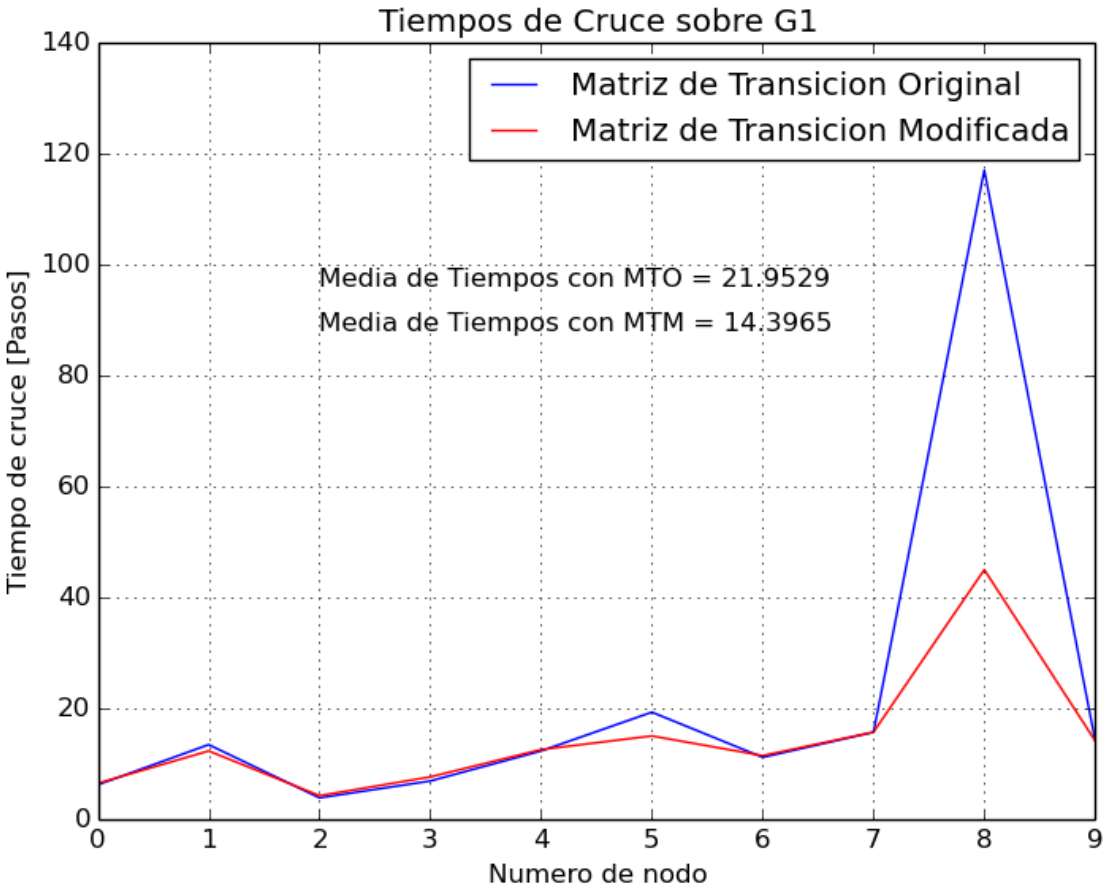


Figura 3: Comparación de los Tiempos de Cruce con las dos Matrices de Transición.

4.2.2. Tiempos de Cruce para G2

Para G2 solo mostraremos la media de los tiempos ya que como dijimos anteriormente, a la hora de mostrar su estructura, posee una gran cantidad de nodos y seria ilegible poner el tiempo de todos ellos.

Pero mostraremos un gráfico que nos ayudara a sacar conclusiones.

Con Matriz de Transición Original:

Media de Tiempos = 103.501

Con Matriz de Transición Modificada (α = 0.85):

Media de Tiempos = 102.522

La Figura 4 nos muestra una comparación de los tiempos de cruce sobre G2 usando las dos Matrices de Transición. Podemos ver que los tiempos de cruce son muy similares y los tiempos medios prácticamente iguales.

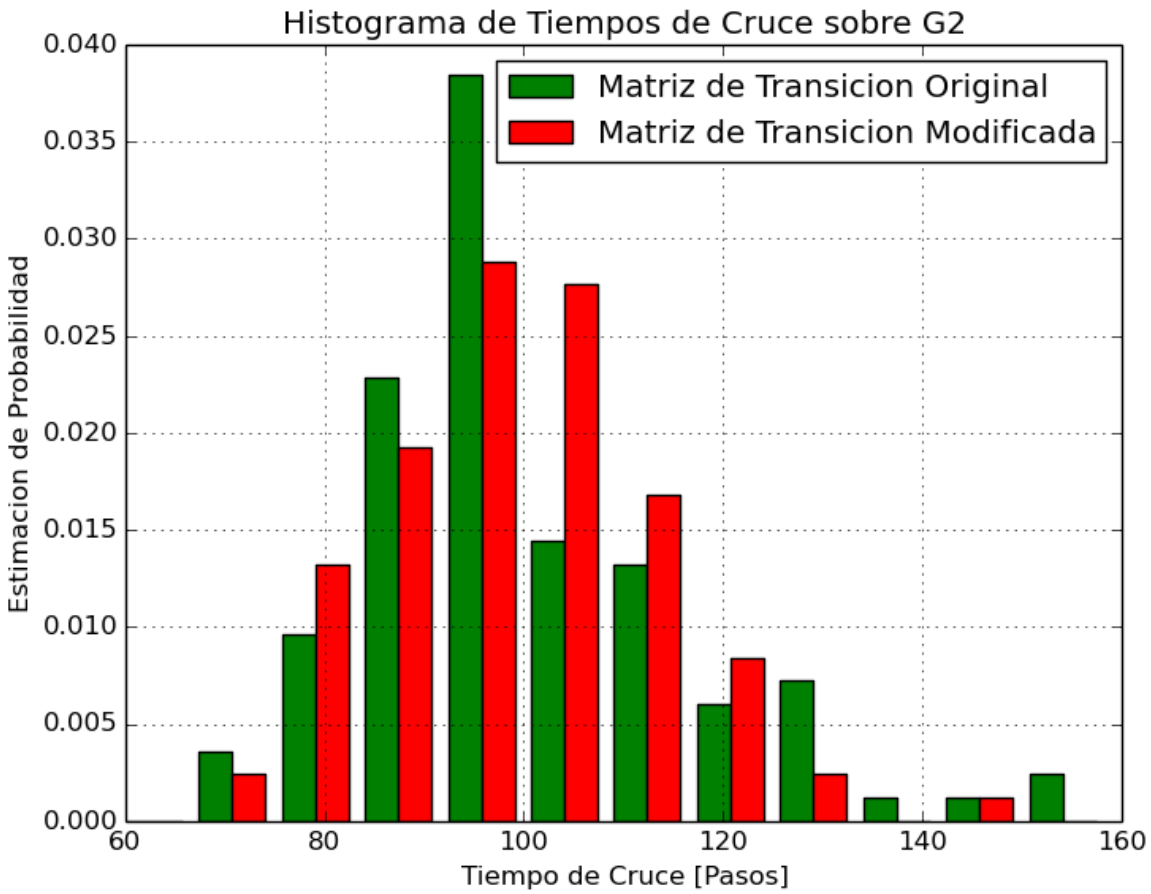


Figura 4: Comparación de los Tiempos de Cruce con las dos Matrices de Transición.

4.3. Tiempo de Cubrimiento

Vamos a calcular el tiempo de cubrimiento sobre G1 y G2 usando las dos Matrices de Transición, pero en la modificada vamos a ir variando el α para ver como varia el tiempo de cubrimiento y compararlos mediante el uso de una tabla.

Tiempos promedio de Cubrimiento con Matriz de Transición Original:

G1	G2
107.641	548.714

Tiempos promedio de Cubrimiento con Matriz de Transición Modificada:

α	G1	G2
0.1	30.449	523.399
0.2	29.942	519.633
0.3	31.549	522.228
0.4	32.85	529.887
0.5	35.213	522.984
0.6	36.81	525.198
0.7	42.012	535.352
0.8	49.959	539.09
0.85	54.777	543.288
0.9	63.136	544.125
0.99	100.901	548.871

En la Figura 5 y 6 podemos apreciar los tiempos de cubrimiento, sobre los grafos G1 y G2 respectivamente, usando las dos matrices de transición a medida que vamos variando α.

Podemos notar en las dos figuras que mientras más grande es α, el tiempo de cubrimiento se acerca más al tiempo obtenido usando la Matriz de Transición Original.

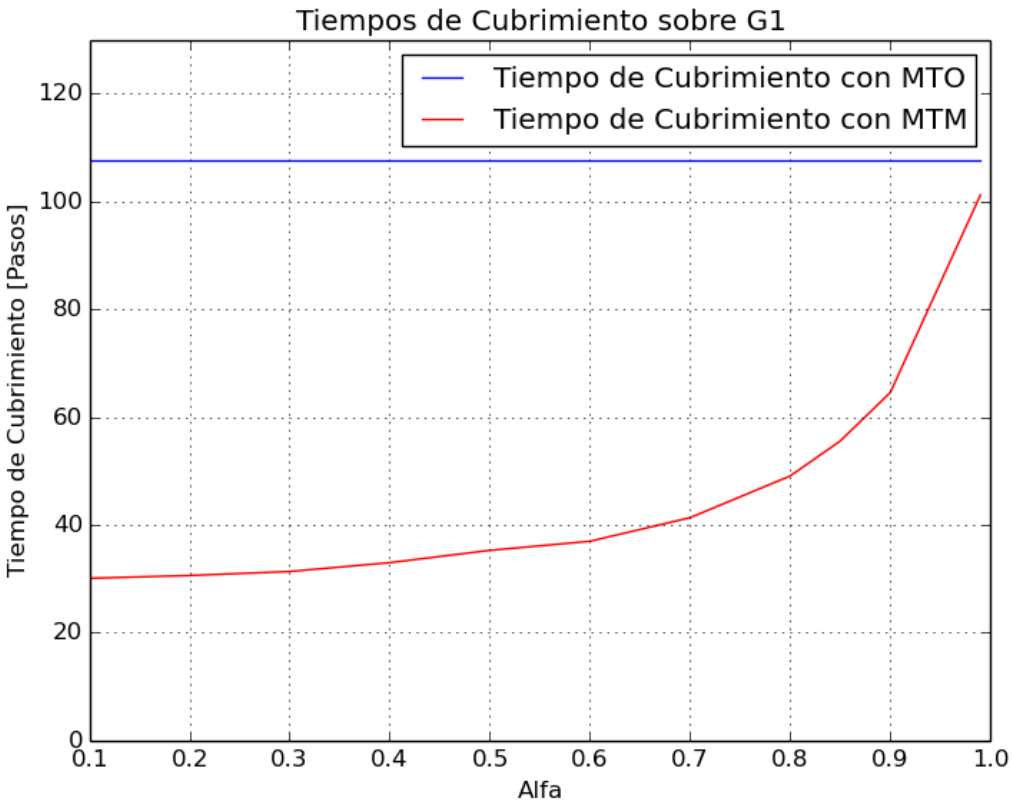


Figura 5: Comparación de los Tiempos de Cubrimiento con las dos Matrices de Transición.

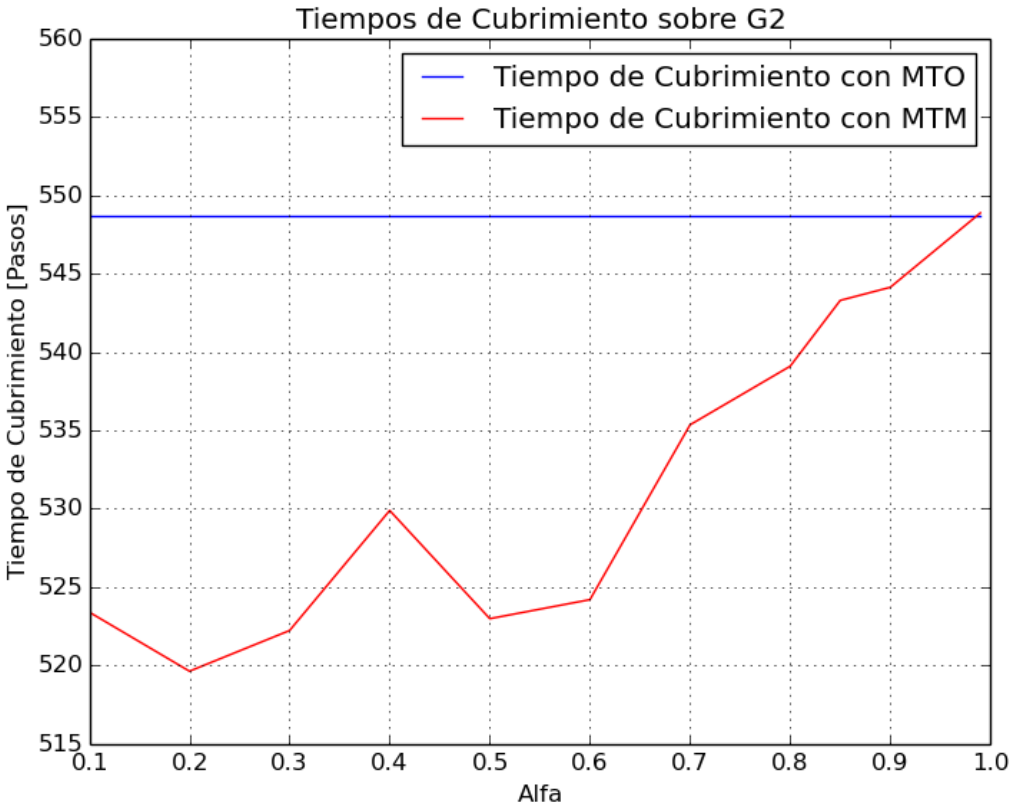


Figura 6: Comparación de los Tiempos de Cubrimiento con las dos Matrices de Transición.

4.3.1. Tiempos de Cubrimiento para grafos de distintos tamaños

A continuación analizaremos los tiempos de cubrimiento al variar la cantidad de nodos de los grafos generados de forma aleatoria, usando la Matriz de Transición Modificada (α = 0.85).

Cantidad de Nodos	Tiempo Promedio de Cubrimiento
5	11.813
10	31.742
30	159.51
50	250.54
75	402.357
100	558.348

Podemos notar mediante la tabla anterior que a medida que aumentamos la cantidad de nodos de un grafo, los tiempos de cubrimiento también aumentan. Esto podemos apreciarlo de una mejor manera mirando la Figura 7 donde vemos que el aumento de tiempo es prácticamente lineal con respecto a la cantidad de nodos.

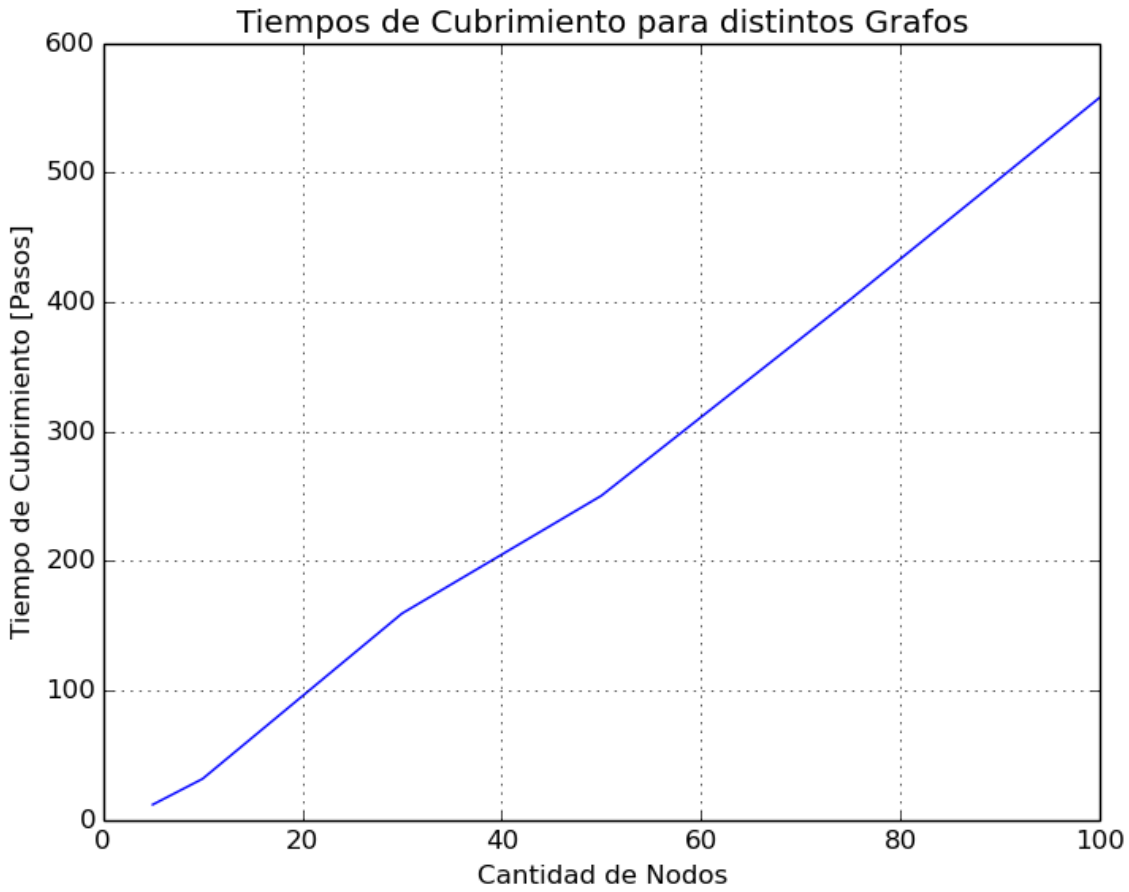


Figura 7: Crecimiento de los Tiempos de Cubrimiento.

4.4. Problema del Spammer

En esta parte vamos a exponer los resultados obtenidos con las dos estrategias que propuso el Spammer para subir el ranking de su página lo que más se pueda.

4.4.1. Estrategias A y B sobre G1

Vamos a analizar dichas estrategias antes mencionadas sobre el grafos G1, comparando como sube o baje el ranking de su página según la cantidad K de páginas ficticias que se creen y hacer que todas ellas apunten a nuestra página, o que se hackeen agregándoles enlaces para que apunten a nuestra página.

Además vamos a comparar los rankings obtenidos usando el Método 2 anteriormente mencionado y la Matriz de Transición Modificada ($\alpha = 0.85$).

K	Ranking con Estrategia A	Ranking con Estrategia B
1	0.023	0.036
3	0.049	0.073
5	0.072	0.12
7	0.09	0.167
10	0.112	0.205

Al analizar la tabla anterior vemos que a medida que aumentamos la cantidad de páginas, el ranking del sitio del Spammer es mucho mejor si se hace uso de la Estrategia B.

En la Figura 8 podemos analizar esto de una mejor manera.

4.4.2. Estrategias A y B sobre G2

Ahora procederemos a hacer lo mismo que en la Sección 4.4.1. pero sobre el grafo G2.

Además vamos a comparar los rankings obtenido usando el Método 2 anteriormente mencionado y la Matriz de Transición Modificada ($\alpha = 0.85$).

K	Ranking con Estrategia A	Ranking con Estrategia B
10	0.014	0.005
25	0.031	0.008
50	0.057	0.017
75	0.08	0.021
100	0.101	0.028

Al analizar la tabla anterior vemos que a medida que aumentamos la cantidad de páginas, el ranking del sitio del Spammer es mucho mejor si se hace uso de la Estrategia A.

En la Figura 9 podemos analizar esto de una mejor manera.

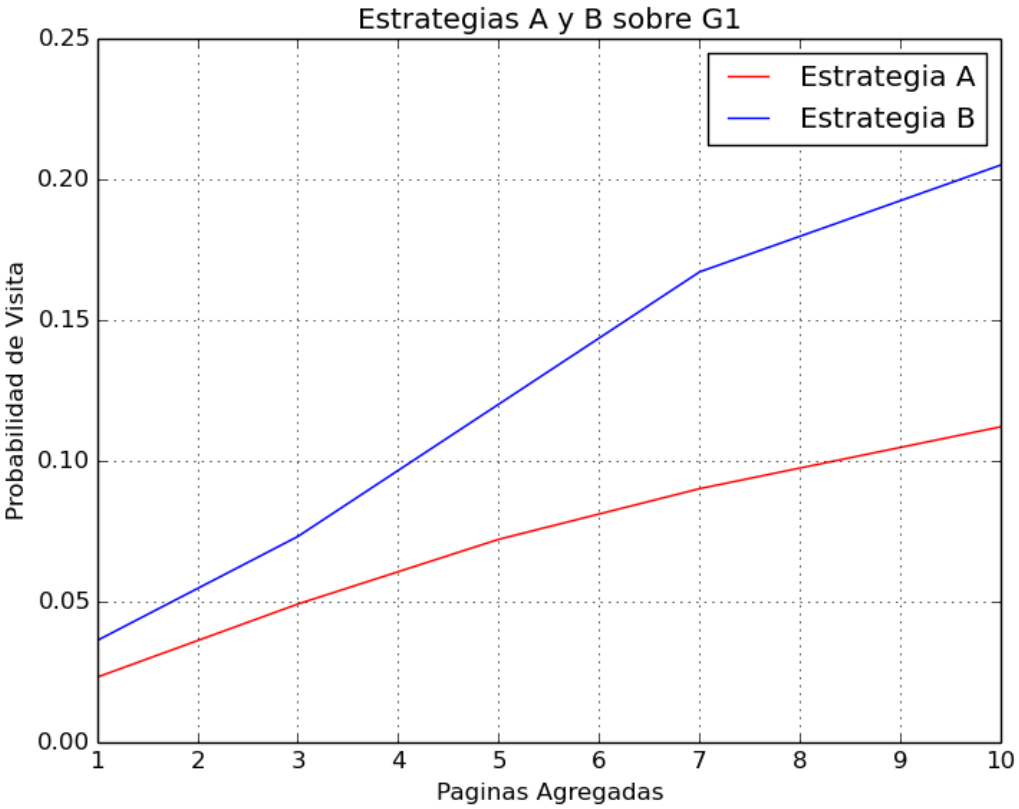


Figura 8: Comparación de Estrategias A y B sobre G1.

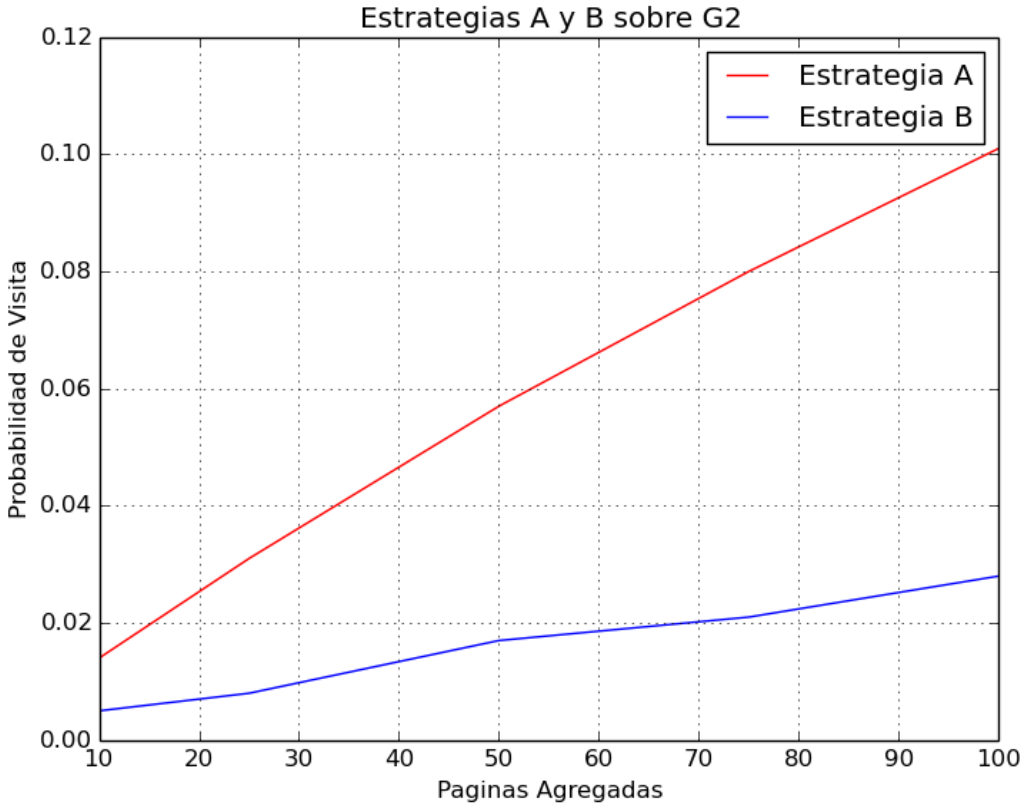


Figura 9: Comparación de Estrategias A y B sobre G2.

5. Conclusiones

A continuación expondremos las conclusiones que hemos obtenido al analizar los resultados y los gráficos mostrados en secciones anteriores:

Distribución Estacionaria

Al calcular la Distribución Estacionaria por medio del Método 1 y 2 con las diferentes matrices de transición, notamos que los resultados son muy similares, por lo que podemos concluir que tanto el método del caminante aleatorio como el método de las potencias son muy buenos para grafos que sean conexos. Por lo que no se tendría que descartar ninguno de los dos métodos a la hora de calcular la distribución estacionaria.

Tiempos de Cruce

Con respecto a los tiempos de cruce de los nodos de un grafo, notamos que al hacer uso de la Matriz de Transición Modificada mejoramos notablemente los tiempos que al hacer uso de la Matriz de Transición Original, frente a grafos cuyos nodos tengan pocos vecinos a los cuales poder ir, es decir, pocos arcos salientes ya que estos generan que el valor medio se vaya a valores altos.

Tiempos de Cubrimiento

Con respecto al tiempo de cubrimiento, mediante un análisis empírico pudimos apreciar que a medida que el α se va acercando a 1, el tiempo de cubrimiento va aumentando acercándose al valor calculado con la Matriz de Transición Original. Esto lo podemos evidenciar mediante las Figuras 4 y 5.

También vimos que si mantenemos un α fijo, el tiempo de cubrimiento del grafo crece prácticamente de forma lineal a medida que se va aumentando la cantidad de nodos del grafo.

Problema del Spammer

Analizando los rankings y gráficos obtenidos en la Sección 4.4. con ambas estrategias y para los dos grafos, llegamos a la conclusión de que para grafos chicos conviene hackear páginas para así poder aumentar el ranking del sitio del Spammer, en contraste con lo sucede en los grafos grandes donde es mejor agregar páginas para aumentar el ranking.

Por lo que concluimos que depende del tamaño del grafo conviene el aplicar una u otra estrategia.